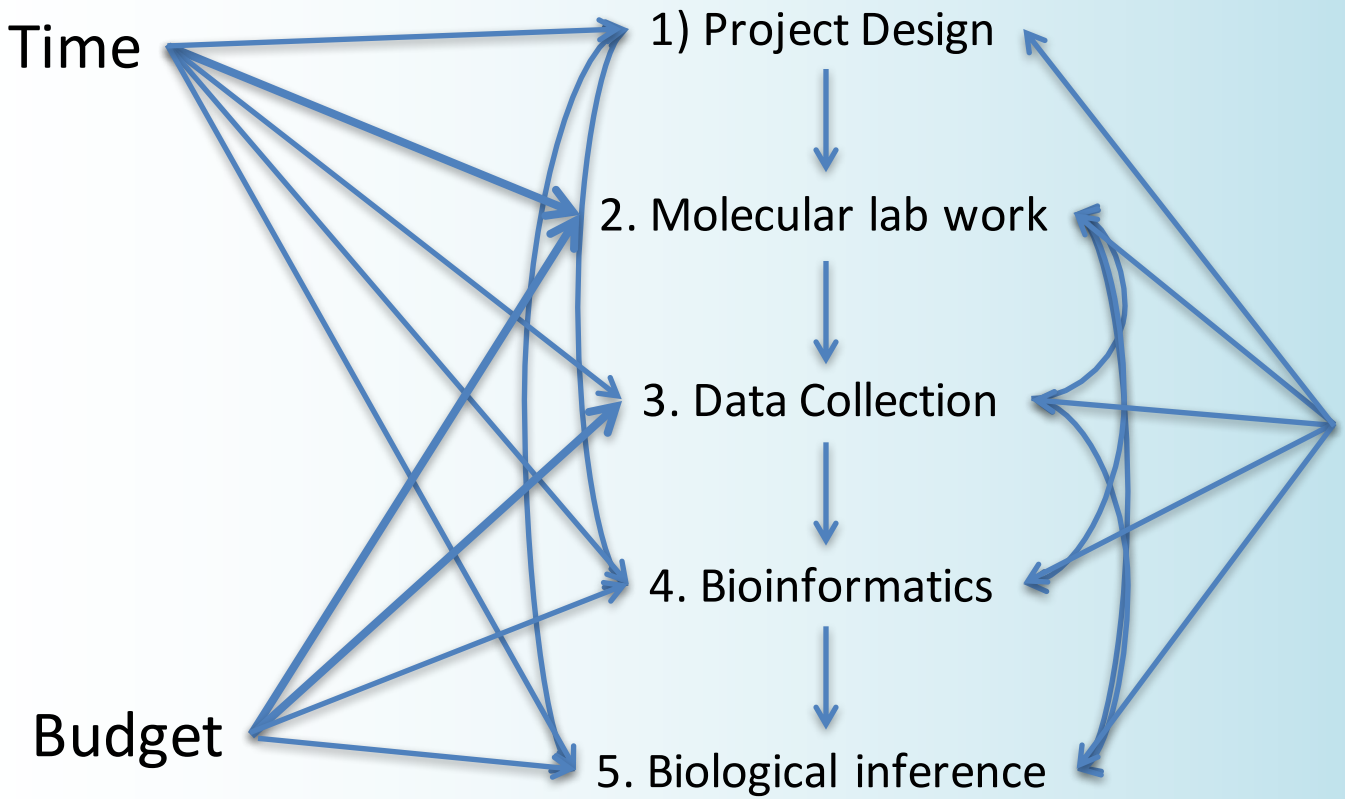
Getting Started in the Molecular Lab with High-Throughput Sequencing: best practices & planning ahead

Lydia Smith
Manager, Evolutionary Genetics Laboratory,
Museum of Vertebrate Zoology
lydsmith@berkeley.edu
28 November 2017

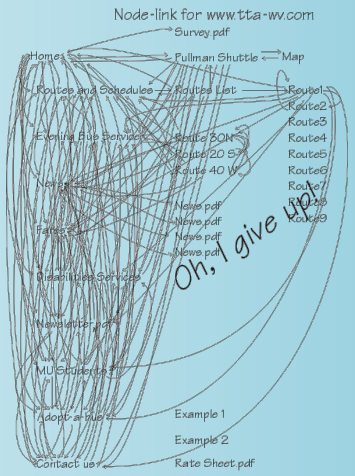
Easy steps to HTS* project success



Other resources (available samples, genomic references, expertise, pipelines, data storage, etc.)

*F.K.A. Next-Generation Sequencing

Horrendogram

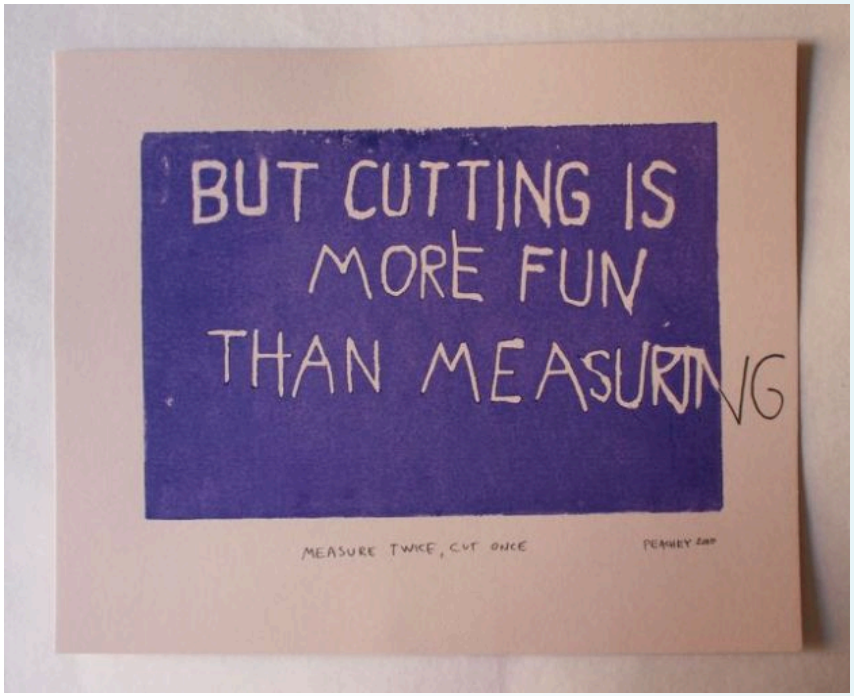


- But don't despair!
- Once you begin thinking about what you really need from your data and what your limitations are in terms of time, budget, and resources, options will disappear and decisions will become easier.
- **The two uppermost parts of the process (Project Design & Molecular Work) determine everything that follows.**
  - Samples can be easily resequenced (although for a high cost) and reanalyzed (although at a cost of time), but if libraries were not made with care, made in the right format, and made to target the most useful type of data, you may not be able to get the data you need from the project without starting over.
- **No bioinformatics tools can save poor data or a poor design**

Problems to consider

- What type of data is best-suited to answer your biologically relevant questions?
- How much data do you need?
- How many samples are needed in the study and available for your research?
  - What the feasibility of getting the sampling you need through fieldwork, museum loans, colonies, etc?
- What is your project budget
- Number of individuals to sequence
- Number of markers to target
- Availability of reference genome
- Coverage needed for downstream analysis

Everything Hinges on Careful Study Design



It is tempting to dive in and just collect a lot of data as quickly as possible, but if these questions aren't thoughtfully considered before starting, the desired bioinformatic approaches and biological inferences may not be possible and large amounts of money can be wasted.

Consult with someone informed as early as possible before you begin lab work: PI, labmate, collaborators, CGRL:

http://qb3.berkeley.edu/cgrl/contacts/

Things to think about in advance of consultation

Choosing a right approach for your project .....

- It depends on your questions:

- *Population genetic parameter estimate (thetas, structure, hybridization, gene flow, change in Ne, etc.)*: RAD, GBS, exon capture
- *Selection on protein evolution*: RNAseq, exon capture
- *Both demography and selection*: exon capture (exons + introns)
- *Other population genetic applications (admixture mapping, constructing linkage map, associations, phylogeography, etc.)*: RAD, GBS, exon capture
- *Larger number (>10,000) of phylogenetic markers*: RNAseq, exon capture, (RAD if shallow divergence)
- *Small number (100s) of phylogenetic markers*: Amplicon sequencing, AHE, SCPP, UCEs

- It depends on the quality and quantity of the DNA

- *Low quality (heavily degraded) DNA*: All hybridization-based methods
- *Low quantity (e.g. a few nanograms)*: nextRAD; certain library prep protocols for sequence captures/RAD

- It depends on the desired sample size (S) + target size (T)

- *Large S + large T*: exon capture, RAD, GBS
- *Large S + small T*: Amplicon sequencing, AHE, UCE, PEC, SCPP, RESTseq

- It depends on your budegt

- *Expensive*: All commercial in-solution exon capture kits
- *Cheap*: All RAD/GBS variants, SCPP, array-based exon capture, UCEs etc.

- It also depends on the genome size/composition, availability of reference resources, the timeline for getting the project finished, experience and support in lab and bioinformatics……

Don’t get overwhelmed: talking about this with an expert will help to narrow down the options and give you particular pathways to focus on

Choosing a right approach for your project ... is beyond the scope of this presentation

Some resources I recommend to new researchers in the EGL trying to start thinking about the techniques we most commonly use:

Genohub's Beginner's Handbook to Next Generation Sequencing
https://genohub.com/next-generation-sequencing-handbook/

Jones & Good, 2016
Targeted capture in evolutionary and ecological genomics. Molecular Ecology
doi: 10.1111/mec.13304

Andrews, et al. 2016
Harnessing the power of RADseq for ecological and evolutionary genomics
doi:10.1038/nrg.2015.28

Think about what biological insights you want, look at what others have done to address similar questions, and then focus your fact-finding on their methods and/or even newer innovations

Researcher Advice: Study Design

- *Start with the biological questions you want to address. I honestly find the most useful thing to find papers / projects that are trying to do something similar and see what approach they used. No one approach is right across projects.*
- *Study design from the very very beginning (when you are deciding how and what material to collect in the field) is really important and it is worth it to spend time talking to as many people as possible before you begin -- i.e. statisticians, lab specialists, and bioinformatics specialists. This will be key to having a successful project. You put a lot of money into library preps and sequencing and get all the data back in one big chunk after investing all this money, so it is difficult to change half way through, although not impossible.*
- *Focus on research questions that expand current knowledge, try to avoid the easy temptation of redoing same the study in a different system. Genomics is just a tool, important and interesting biological questions are what students should spend most of their time thinking about, especially early on.*
- *Overshoot the samples needed to account for failures in sequencing, if possible. You can have an ideal study design compromised from a failed sample or two and resequencing isn't always very easy.*
- *The NGS world changes extremely fast, the best strategy is to pick the best method at the moment you start a project and then move forward. There always be somehting better.*
  - We can't get too hung up on trying to do things the optimal way or else we will never actually get started
- *If you are doing something new, you won't necessarily get it perfect the first time. But it is still worth doing since neither the methods or the science will move forward unless people take the risk. A project should have multiple levels so that if the data can't answer the more ambitious question, it can still be used for other meaningful (and publishable) biological inferences.*

Campus Resources for high-throughput sequencing: QB3 Labs

- Computational Genomics Resource Laboratory (CGRL)
  - project design & preliminary bioinformatics discussions prior to start of a project
  - **talk with them before beginning library preparations**
  - computer cluster and pipelines designed for large-scale genomic data
- Functional Genomics Laboratory (FGL)
  - will perform library preparations in conjunction with the GSL. This is more expensive per prep than DIY approaches (~$75 genomic DNA, ~$150 RNA with Poly-A selection) but a great option for projects with very large budgets or only a few libraries to make.
  - has specialized pay-per-use equipment for researchers making their own libraries: Covaris, Bioanalyzer, Pippin Prep
- Vincent Coates Genomics Sequencing Laboratory (GSL)
  - supports Illumina and PacBio sequencing technology on campus
  - request to be subscribed to their user list by emailing gslusers-owner@lists.berkeley.edu.

Getting the best possible data from HTS

HTS has a very high buy-in cost and no one wants to get stuck with a giant pile (in the case of Illumina: 100's of millions!) of sub-par reads (or of great reads that can't answer our scientific questions. It pays in the long run to take the time to do the best possible job handling DNA and libraries in the molecular lab:

- Have a clear plan for all your molecular lab decisions before beginning any lab work
- Budget enough time to work cautiously and to be able to repeat samples when needed.
- Start with high-quality DNA extractions
- If making libraries yourself, sweat the small stuff. Take time to get the sizing correct before proceeding to enzymatic steps
- If using a paid provider, be sure that they are reputable and produce good data, and be sure to effectively communicate all sample processing considerations to them (don't make assumptions)
- Use a well-regarded sequencing facility, <u>not</u> the cheapest option you can find

Getting the best possible data from HTS

Getting the best possible data from HTS

- Yes, it is expensive
- Yes, it is time-consuming
- Yes, it is confusing
- Yes, it can be stressful
- But you can also enjoy the benefits of new technologies revolutionizing the types of biological questions that can now be addressed (relatively) cheaply even in non-model organisms
  - https://doi.org/10.1016/j.margen.2016.04.012
- HTS is also more robust than traditional sequencing workflows
  - It is not the same concept as setting up 1000's of Sanger sequencing reactions at once and crossing our fingers that everything works and we don't waste our entire budget
  - Numerous quality control steps allow us to weed out or replace dud samples prior to sequencing
  - I haven't heard of any failed projects that didn't ignore big warning signs from QC: all eggs may end up in the same basket, but it is a *really* well-constructed basket.

About the Evolutionary Genetics Laboratory

- Molecular biology facility for the Museum of Vertebrate Zoology (and friends)
- Since we are limited by funding, reference genomes, and computational resources, we usually can't sequence the whole genome of any organism of interest.
- Instead, like most museum labs, we must find ways to:

  1) reduce the amount of sequence data collected per sample, while

  2) ensuring that we obtain orthologous regions of the genome from all the samples, and

  3) that we collect the type of genomic data that allows us to address biologically meaningful questions, while

  4) keeping costs as low as possible.

About the Evolutionary Genetics Laboratory: How we work

- We are not a fee-for-service lab like the ones previously mentioned. Instead we offer training and trouble-shooting support, access to specialized equipment (bioruptor sonicator, plate magnets, hybridization equipment), bench space, and a storeroom with essential reagents and adapters in stock in order for researchers to perform their own wet-lab work.
- We are the molecular biology facility for the Museum of Vertebrate Zoology, but we also allow lab membership for anyone in the wider Berkeley community so long as they are:
  - working on project types that we support (especially, but not limited to, genomic DNA capture by hybridization, transcriptome/RNA-Seq, RADseq);
  - willing to put up with some of the challenges of a shared lab environment as well as reaping the benefits (*patience* is key) and to act as a member of a lab community;
  - able to complete the full orientation and safety training for our facility;
  - in agreement to purchase the bulk of their project supplies through our storeroom (the overhead percentage funds equipment, and salaries; for common items, our discount for bulk purchases is greater than the overhead.)

About the Evolutionary Genetics Laboratory: Supported Techniques

Fully supported (all supplies in stock except custom hybridization probes) for DIY researchers:

- Genomic DNA library preparation with PCR based on Meyer & Kircher 2010 (reagents, plastic consumables & QC = $18-23/library)
  - Historical libraries with USER-treatment = $22-27/library
- Genomic DNA library, PCR-free (kit, dual-index adapters, plastic consumables & QC = $30-45/library)
- ddRAD-seq library based on Peterson, et al. 2012 (library reagents, plastic consumables, size-selection & QC = $10-15/library)
- RNA library, poly-A selection (kit, dual-index adapters, plastic consumables & QC = $30-$45/library)
- Hybridization captures with Nimblegen or MyBaits probes (buffers, Cot-1, blocking oligos = $120/capture)

Partially supported (can work with you to get additional items as needed)

- RNA library, rRNA depletion
- Amplicon sequencing

About the Evolutionary Genetics Laboratory: Project Types

In the past two years:

Capture Hybridization methods:

- Agilent chips: spiders, plants
- Nimblegen SeqCap in-solution: jaguars, rodents, cone snails, hummingbirds
- MyBaits in-solution (custom): frogs, lizards, plants
- MyBaits UCE's: birds, lizards, ants
- RNA-Seq: rodents, bats, plants
- WGS: birds, golden toad, nematode, plants, foxes
- ddRAD: plants, molluscs, rodents, beetles, birds, penguins
- hyRAD: armadillo

About the Evolutionary Genetics Laboratory: Specialized Equipment

- Diagenode Bioruptor (sonicator)*
- Qubit (fluorometer)**
- Bioanalyzer**
- Magnetic plates
- Hybridization oven for array-based captures
- MoBio PowerLyzer (bead beater for RNA homogenization or difficult tissue DNA extractions)

*Covaris sonicator available in FGL

**Qubit & Bioanalyzer available in FGL & GSL

About the Evolutionary Genetics Laboratory: How to get involved:

- E-mail me with a brief description of your project, your past experience, and your time-table for completing the work: lydsmith@berkeley.edu
- We will discuss whether your project is a good fit for our resources, expertise, and availability
- We can schedule orientation and training sessions when time allows
- Payment for storeroom purchases: UC Berkeley chartstring or invoice information if funding is off-campus
- We have only one employee (that's me!), so I can't offer consultation to people who are not members of our lab or who are not purchasing from our storeroom

Types of High-Throughput Sequencing Platforms and Methods

Short-read:

- Illumina single-read or paired-end
  - short reads up to 300 bp, huge market share, best cost per Mb

Long Read/Single Molecule:

- Pacific Biotechnologies
  - reads ~10 kb, whole genome sequencing, full-length cDNA transcripts
- Oxford Nanopore Technologies
  - possibility of ultra-long reads (> 100kb)

"Synthetic" Long Reads:

- 10X Genomics Chromium
  - Partitions and internally barcodes long DNA molecules (50-100 kb) that can be bioinformatically assembled into "synthetic" long reads. Requires special machinery for library preparation (UC Davis) but can be run on a regular Illumina sequencer
  - Similar Chromium technology available at Berkeley GSL for single-cell RNA-Seq

Scaffolding:

- HiC/CHiCago
- Mate-pair libraries

For more options and details, see: CGRL Genome Assembly Workshop:
http://cgrlucb.wikispaces.com/file/view/Assembly_Workshop_Berkeley_2017.pdf/618591217/Assembly_Workshop_Berkeley_2017.pdf

How to chose a HTS Platform

- What machines are available? (on-campus, other UC, reputable universities/companies)
- Do you need long reads? How long?
- Do you need paired-end reads?
- Do you need high coverage?
- Do you need large-scale multiplexing?
- What genomic resources do you have for these organisms?
- Budget

Sequencing Machines 2017
UC Berkeley-supported technologies

*Illumina buy-in costs do not include library prep since that can vary from insignificant (< $100) to far more than the cost of the instrument run

Machine	Best Cost ($/GB)	Run buy-in cost	Reads per run (millions)	Read Length	Paired-end?	Multiplexing?	Final Error Rate
ABI 3730 Sanger sequencing	$1.5 million	$4	0.000001	Up to 1000 bp	yes	no	0.1-1%
Illumina HiSeq 4000	$20	$1100-$2450*	350+ (SR) per lane	Up to 150 bp	yes	yes	0.1%
Illumina HiSeq 2500 (Rapid)	$46	$2500-$7420*	240+ (SR) per flowcell	Up to 250 bp	yes	yes	0.1%
Illumina MiSeq	$145	$1100-$1750*	15-25 (SR) per lane	Up to 300 bp	yes	yes	0.1%
PacBio Sequel	$375	$1500	0.385	Up to 10kb	no	yes	≤1%

http://www.molecularecologist.com/next-gen-fieldguide-2016/

Why so many types of Illumina instruments?

- HiSeq 4000: workhorse for short reads (up to 150bp)
  - Most cost-effective choice for most large projects using short-read data
  - Index swapping issue makes it less well-suited for low-coverage data (unless unique dual indexes are used).
- HiSeq 2500: up to 250bp reads
  - Best approach when slightly longer reads are needed (250PE)
  - Good choice for some unusual library distributions
  - Does not display evidence of index swaps
- MiSeq: small projects, up to 300bp reads
  - Sometimes only a small amount of data is needed to answer a research question
  - Preliminary runs to test library pools before investing in bigger sequencing lanes
  - Best approach for amplicon sequencing that need longer reads (up to 300 PE) in order to reach the middle of the library and that have low diversity

Additional Illumina Options

- MiSeq nano (Berkeley GSL):
  - 1 million reads, PE150 = $400
  - Can be used as a QC check before embarking on large sequencing efforts to ensure that libraries and pooling are good
- HiSeq X (Hudson Alpha):
  - 400 million reads, PE150 = $1650.
  - High-coverage (~30x) whole-genome sequencing only
- NovaSeq 6000 S4 (UCSF in 2018):
  - 2.5 billion reads (!!!), PE150 = $8000
  - No restrictions on type of sequencing
  - But due to higher error rate of the two-color technology, I'd not recommend it for low-coverage projects

Illumina technology introduction:
https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Local UC Sequencing facilities

QB3 Vincent Coates Genome Sequencing Laboratory: vcgsl.qb3@gmail.com
http://qb3.berkeley.edu/gsl/rates/ *
Director, Shana McDevitt: shana.mcdevitt@berkeley.edu

- Illumina HiSeq4000, HiSeq2500 (Rapid), MiSeq
- PacBio Sequel

UC Davis Genome Center: dnatech@ucdavis.edu
http://dnatech.genomecenter.ucdavis.edu/prices/ *
Director, Lutz Froenicke: lfroenicke@ucdavis.edu

- Illumina HiSeq4000, NextSeq, MiSeq
- PacBio RSII and Sequel
- **10X Genomics Chromium

UCSF Center for Advanced Technologies
https://sites.google.com/site/sequencingucsf/rates *
Director, Eric Chow: eric.chow@ucsf.edu

- **Illumina NovaSeq 6000 S4 coming early 2018!

*All rates subject to change without notice
**Submissions for these instruments can be coordinated through the Berkeley GSL

So, what kinds of data are we able to collect with these instruments?

- Most high-throughput sequencing discussion by vendors and in the media focus on human genetics or other model organisms with medical implications
- $1000 genome hype: depending on the completeness of the assembly, the coverage required, and the calculation of bioinformatics costs, we are already there for re-sequencing
- But for comparative projects with large sample sizes, even that price can be prohibitive
- And de novo sequencing of large & complex genomes still costs $10,000+
- However, sometimes a genome must first be sequenced in order to obtain the information needed for comparative methods.

Whole Genome Sequencing (WGS): De novo vs re-sequencing

- Very different approaches with the same general name (WGS)
- Re-sequencing a genome can be done at a relatively low cost using short-read data alone **if** there is a high-quality well-assembled genome
- Illumina paired-end data is almost always insufficient alone for a high-quality de novo genome. Must be scaffolded with long reads (PacBio, Oxford nanopore), synthetic long reads (10x Chromium), mate-paired libraries and/or chromosome-level information from HiC/CHiCago
- Great two-part talk on de novo WGS by Stefan Prost:
  - http://cgrlucb.wikispaces.com/file/view/Assembly_Workshop_Berkeley_2017.pdf/618591217/Assembly_Workshop_Berkeley_2017.pdf
- Beautiful de novo poster from former EGL undergrad Josh Peñalba:
  - https://berkeley.box.com/v/Penalba-CBA-GenomeSequencing
- Will not discuss de novo much here because the researcher hands-on work in the molecular lab is limited in most cases to DNA extraction
- That said, all of the non-Illumina methods require DNA that is high quality and high molecular weight; molecular lab work is minimal but critical

Whole Genome Sequencing

- A single genome or transcriptome sequence is rarely the end goal of a scientific project.
- Most projects require comparisons between multiple samples to answer the biological questions of interest.
- This can be accomplished with WGS if:
  - there is already a well-assembled reference genome, and
  - the genome size is small, and/or
  - only low levels of coverage are required
- Sequencing whole genomes has some drawbacks (cost, data storage, complexities of analysis), but it allows researchers to look comprehensively at the entire genome, including promotors, regulatory elements, and introns, not just the small slice of the genome in the other methods we will discuss.
- When looking for signatures of selection and patterns of molecular evolution, a WGS will give the study a much greater chance of success than exome sequencing or RAD-Seq alone.
- Even when genomic resources exist and coverage can be low, only a few large genomes can be run on a single lane so sequencing costs can be expensive.

Comparative Genomics: genome partitioning methods

If the size of the genome is small, whole genome sequencing of many samples is a real possibility to consider.

Otherwise, we must find ways to:

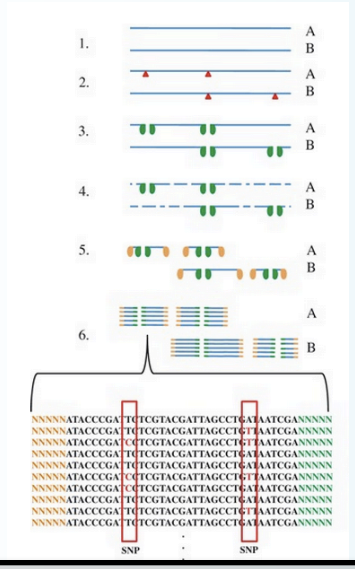
1) reduce the amount of sequence collected while
2) obtaining orthologous parts of the genome from all the samples and
3) ensuring that we collect the type of genomic data to allow us to address biologically meaningful questions

Common methods involve different types of library preparations:

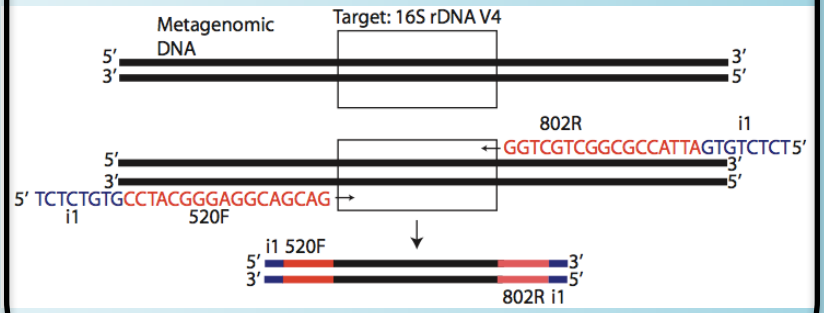
- Targeted capture of genomic DNA (often exons, but could be any known region of the genome)
- Transcriptome sequencing (sequences of exons, differential gene expression)
- RAD-Seq (sequences of SNPs near restriction enzyme cut sites)
- ChIP-Seq (sequences of DNA at and near chromatin protein binding sites)
- Amplicon sequencing (sequences of multiple organisms in one or more PCR amplicons: metagenetics)

Comparative Genomics Tools: strategies depend on the research questions

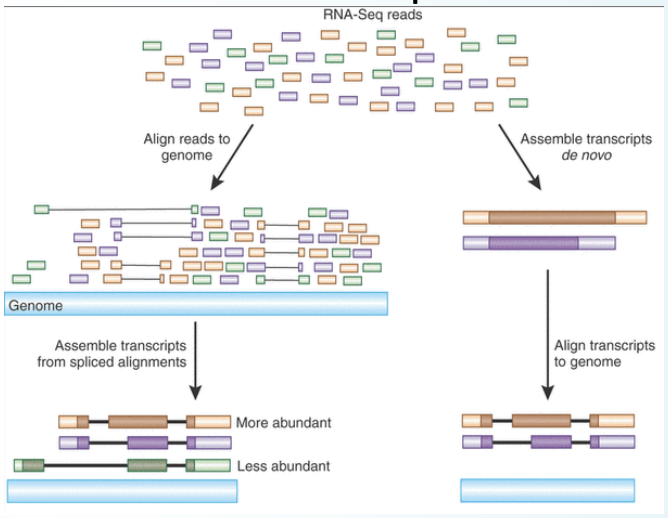
RAD Sequencing



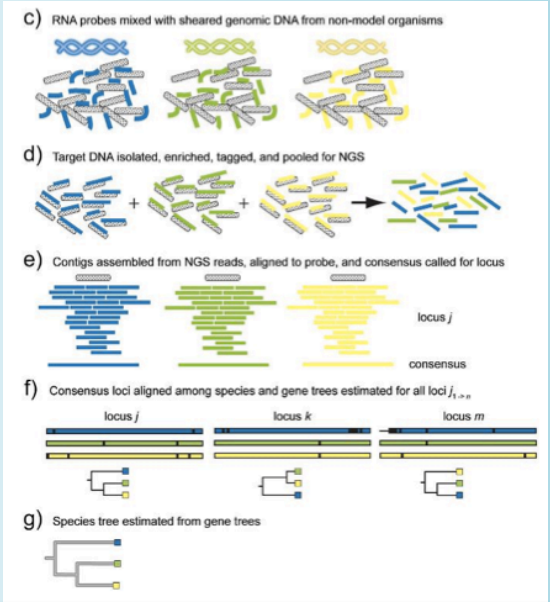
Amplicon Sequencing



RNA Seq



Targeted Enrichment



Targeted enrichment by capture

- Great approach for getting consistent sequencing results from the same genomic partition from multiple samples for both population genetic & phylogenetic studies
- Can multiplex dozens of samples per lane, depending on target size
- Can be used with degraded samples
- Can target specific high-value genomic regions

- Requires a reference genome or transcriptome: *a priori* knowledge of target sequences in order to design custom probes
  - Some predesigned kits are available). If unavailable, preliminary sequencing costs of transcriptome(s) or low coverage WGS will be additional cost/time
- Potentially expensive projects: library preparation costs are more than with RAD-Seq; additional probe synthesis and capture costs
- Challenging for organisms with large genome sizes and/or highly repetitive genomes

Targeted capture in evolutionary and ecological genomics
Molecular Ecology 2015: DOI: 10.1111/mec.13304

RNA-Seq

- Sequences the expressed transcripts in a tissue
- Data has manypossible uses: can be used to get exon sequence information, transcript counts for differential gene expression studies, and/or alternative splicing information
- No previous genome or transcriptome information required
- Requires properly-preserved tissue to extract high-quality RNA
- More expensive library preparation than DNA methods
- Highly expressed genes can dominate the data; deep coverage required to find rare transcripts

Project design and workflow references:

- https://rnaseq.uoregon.edu/
- http://sfg.stanford.edu/guide.html
- https://bioshare.bioinformatics.ucdavis.edu/bioshare/download/kveirzo6fvkl2nb/Thursday_MS_RNASeq.pdf
- https://angus.readthedocs.io/en/2014/_static/MegStaton_NGS_KBS_Staton_RNASeq.pdf

Restriction-site Associated DNA markers: RAD-Seq

- Inexpensive way of identifying 1000's of SNPs in many closely-related organisms. Can gather large number of anonymous markers from many samples over a short period of time
- Widely used for inferring population structures, phylogeography, trait mapping, genetic maps, and association – plenty of case studies in the literature
- No pre-existing reference sequence needed (but having one helps a lot)

- Requires high quality DNA
  - some hybrid approaches like HyRAD (https://doi.org/10.1371/journal.pone.0151651) and Rapture (https://doi.org/10.1534/genetics.115.183665) may allow integration of degraded samples
- Collecting homologous markers from all samples is a struggle no matter how careful the planning and the lab work: problems with locus drop-outs and high variance of depth across loci and individuals
- Poor choice for deep-scale phylogenetics due to potential of mutations at RE cut-sites
- Challenging for organisms with large genome sizes and/or highly repetitive genomes

Harnessing the power of RADseq for ecological and evolutionary genomics
Nature Reviews Genetics 2016: doi:10.1038/nrg.2015.28

Amplicon Sequencing

- Useful for metagenomics applications where there are multiple taxa present in each sample
  - e.g. 16S for investigating diversity and structure of complex microbial communities & populations)
- Cost-effective if # of loci is low and # of samples is high
- No reference genome or transcriptome information needed; just enough sequence to design primers from

- Not cost or time-effective for many loci unless emulsion PCR is used
- Requires a successful amplification for each sample
- Amplicon length is limited by sequencing read lengths
  - since the libraries are not sheared, information more than 300bp from each end cannot be sequenced with Illumina technology
- Environmental contamination is very difficult to avoid
- Low complexity libraries can pose sequencing challenges

How to get from nucleic acid to sequencing

- A library must be first be constructed
- Not an insignificant cost: if multiple samples are run in the same lane (multiplexing), sometimes library preparation costs exceed run costs
- QB3's GSL/FGL will prepare libraries starting at $50 (amplicons), $75 (gDNA), $150 (RNA) : a good option to consider if you have a large budget, few libraries, and/or little available hands-on time:
  - http://qb3.berkeley.edu/gsl/rates/
  - Provide most common Illumina & PacBio library types except for RAD-Seq
- High-quality outside providers include (but are not limited to): Arbor/Mycroarray (gDNA library prep, captures), Floragenex & SNPsaurus (RAD-Seq), Argonne (16S Amplicon)
  - But be very cautious with selecting an outside provider, since many (esp. for-profits) cut corners and produce sub-par libraries
- Otherwise, purchasing a kit or using the EGL can bring costs done to ~$10 (ddRAD)-$50 (RNA), but you have to put the labor in yourself

What is a library?

Libraries take genetic material and make it sequenceable by adding short oligos with known sequences called adapters to the ends of DNA molecules

Often library preparations also involve:

- Fragmenting DNA and size-selecting for particular DNA lengths
- Repairing DNA damage
- Reducing the portion of the genome available to be sequenced

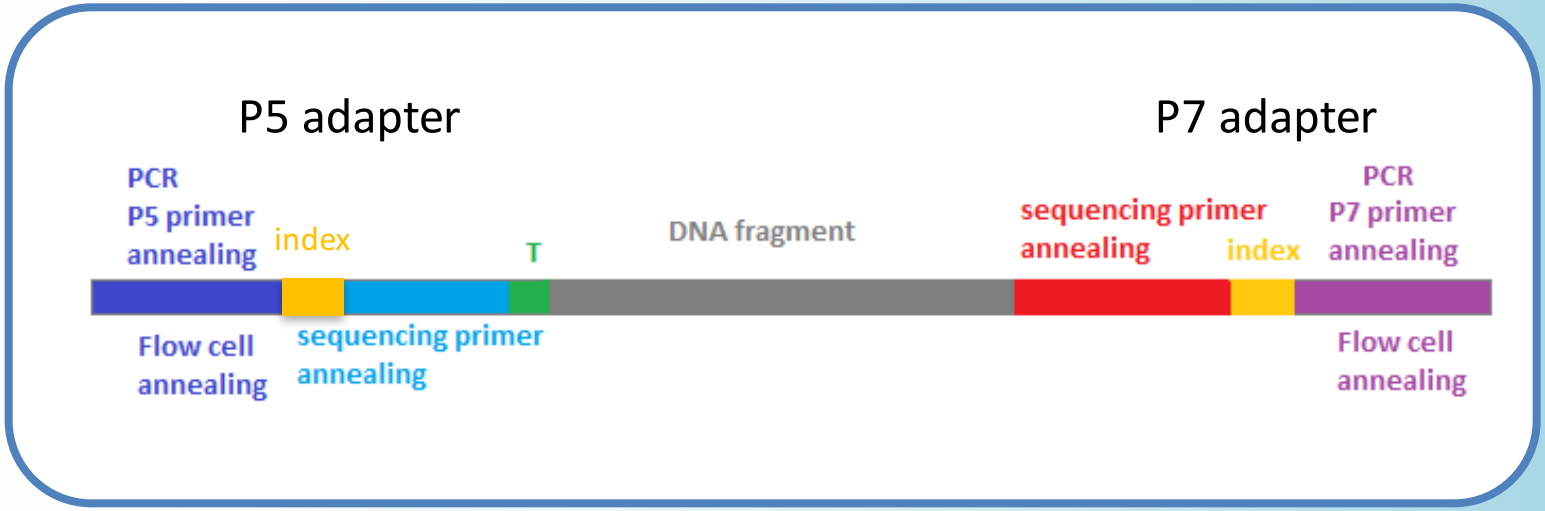
What is a library? (Illumina)

- DNA is fragmented to < 1000bp: sonication or enzymatic reaction.
  - If starting with RNA, it is converted to cDNA before library prep. All Illumina libraries are DNA libraries and must be double-stranded
- Ends are repaired to be fully double-stranded and are prepared for ligation with phosphorylation and/or A-tailing
- Adapters are ligated onto the ends
- PCR enrichment amplifies the number of complete library molecules (optional, depending on method)
- Although all Illumina library preparation methods can be run on all Illumina platforms, some types are much better fits on certain platforms.
  - Know your expected platform before beginning libraries

Anatomy of an Illumina Adapter

An adapter molecule has three parts:

1) Flow cell annealing site (outer adapter)
2) External index (read separately and linked to sequence reads in the metadata)
3) Sequencing primer annealing site (inner adapter)



No matter how we make an Illumina library, we have to end up with all three elements attached to our DNA fragments

Anatomy of Illumina Data

Paired-end, dual-index library:

Paired-end, dual-index data:

READ 1 DATA

```
@M02326:36:000000000-D2U3W:1:1102:12515:2797 1:N:0:TCGCAGG+TCGCAGG
GCCGATCTCCTGGTCCATGCCGCGGATCGCGCCGTCGTAGAGGTCCTTGTTGTGCGCGACGTAGGCGCTCGGGA…
+
BBBBBBBFFFFFGGGGGGGGGGGGGGGGGGGGGGGGGGGEHGGHHGHHHHHHHHHHGHGGGGGGGHGGHGGGGGGGG…
```

READ 2 DATA

```
@M02326:36:000000000-D2U3W:1:1102:12515:2797 2:N:0:TCGCAGG+TCGCAGG
CCCGAGCGCCTACGTCGCGCACAACAAGGACCTCTACGACGGCGCGATCCGCGGCATGGACCAGGAGATCGGCA…
+
ACCBCBCCCCBCGGGGGGGGGGGGHGHHHHHHHHHHHHGGGGGGGGGGGGGGHGGGGGGGHHHHHHHGHGHHHHGGG…
```

https://en.wikipedia.org/wiki/FASTQ_format
https://www.drive5.com/usearch/manual/quality_score.html

DNA/RNA Preparation

- **For most projects, the extraction, assessment, and fragmentation of nucleic acids will by far be the most time-consuming part of the lab work.**
- Sometimes this may feel frustrating—like you are spinning your wheels before getting started on the real work—but DNA/RNA preparation and quality assessment is the most important part of the library preparation process.
- Nothing is more correlated with success than starting with sufficient quality, quantity, and sizing of nucleic acid for your project needs.
- The best quality DNA/RNA is going to give us the best sequencing results

Nucleic Acid Prep *is* Library Prep

- For most projects and library protocols, the most important choices and actions are made at this stage
  - How much input material
  - How long of an insert size to aim for
  - Assessment to ensure that the proper length was obtained
  - Use of size-selection to refine overall insert size
- Often samples behave slightly differently from one another and have to be assessed and handled individually
- But once all samples have been processed to be the desired concentration, size, and volume, they can usually be treated the same throughout the enzymatic steps

DNA Extraction

1) Take care to minimize contamination since every double-stranded molecule inside your sample can be turned into a library.
2) For better assessment, incorporate an RNaseA digestion step into your DNA extraction
3) Avoid spin column kits and centrifugation steps if high molecular weight DNA is required
    - See Appendix for recommendations and protocols
4) Elute/Resuspend your DNA in a buffer with low or no EDTA buffer: 1x LTE (10mM Tris, 0.1mM EDTA) or Qiagen EB (10mM Tris pH 8.5)

DNA Assessment

1) Qubit: dye-based assessment (fluorometer) which will can specifically assess double-stranded DNA
   - spectrophotometer readings (ie. Nanodrop) will collect data from any nucleic acid (or anything else absorbing light at 260 nM), even mononucleotides!
   - But almost all DNA library processes will only work with dsDNA

The Qubit uses reagents, so there is a charge per use. However, it is worth the money to get accurate **dsDNA** information since that is the only material which will get made into libraries.

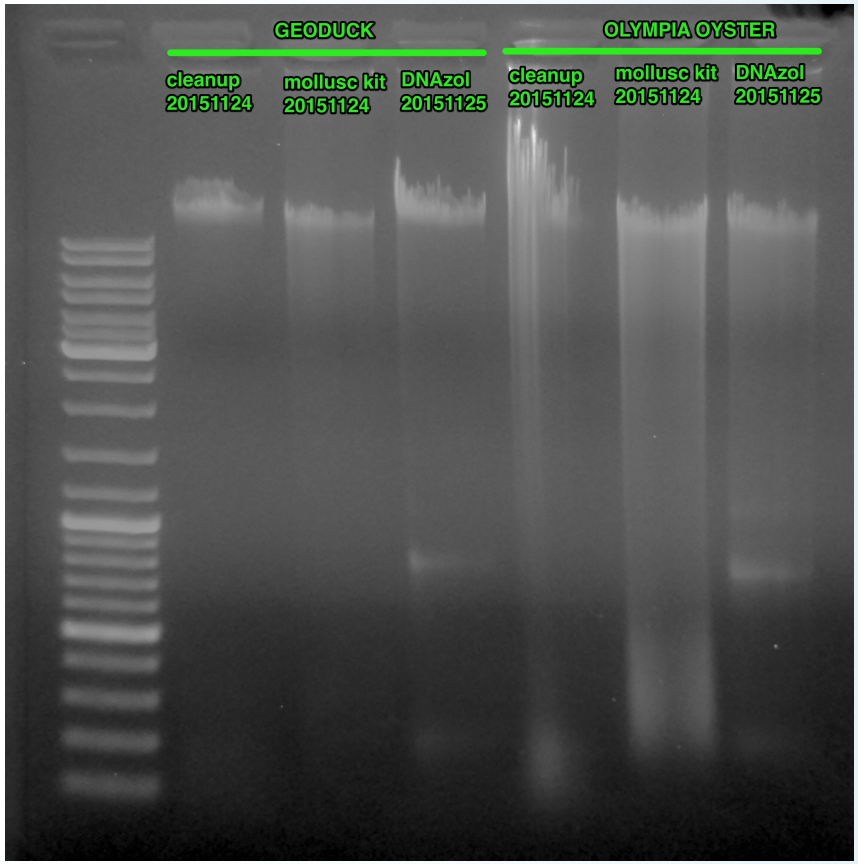
FGL/GSL: $4.00/sample

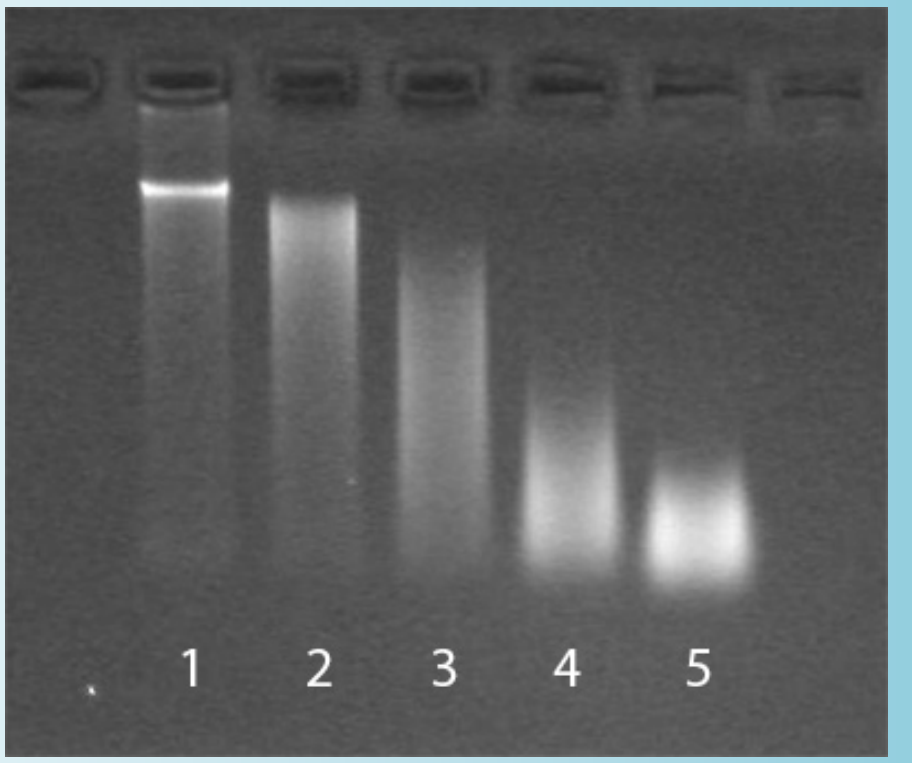
also available in the EGL for members of that lab

DNA Assessment

2) Run DNA samples on an agarose gel to check for a high molecular weight band and signs of any degraded sample/residual RNA.

Use fresh (non-recycled) running buffer; post-stain in bath if possible





Note: degraded samples <u>can</u> still be used for many library preps, but must be handled differently

DNA Assessment

3) Microfluidics devices: Bioanalyzer, Fragment Analyzer

- Electrophoresis instrument performs size fractionation and quantification of small samples of DNA, RNA, or proteins.
- Provides much more accurate sizing information than agarose gel electrophoresis
  - Concentration data not considered as accurate as Qubit
  - It is also redundant if you already have sizing data from agarose gel
  - However, can be useful for getting a more detailed and accurate look at the distribution of fragments of degraded samples.
  - Bioanalyzer traces are essential at the end of the library prep process for accurate quality assessment

FGL/GSL: $10.00/sample

also available in the EGL for members of that lab

RNA Extraction

Caution: this is not meant to be an exhaustive list. RNA is far pickier than DNA. If you haven't worked with it before, please consult others with experience for their detailed advice before embarking on collection of tissues meant for RNA research.

1) Proper handling of tissue is key: a) flash-frozen or b) RNA Later overnight then stored at -80°C or below. Thawed as few times as possible.

2) Be very cautious to avoid environmental RNAses during the extraction process (as well as post-extraction handling): use RNA-restricted equipment & plastics, clean well with Eliminase/RNA-Away, change gloves as often as needed…

3) Homogenize samples well. If possible, use a mixer or bead beater to homogenize rapidly and thoroughly

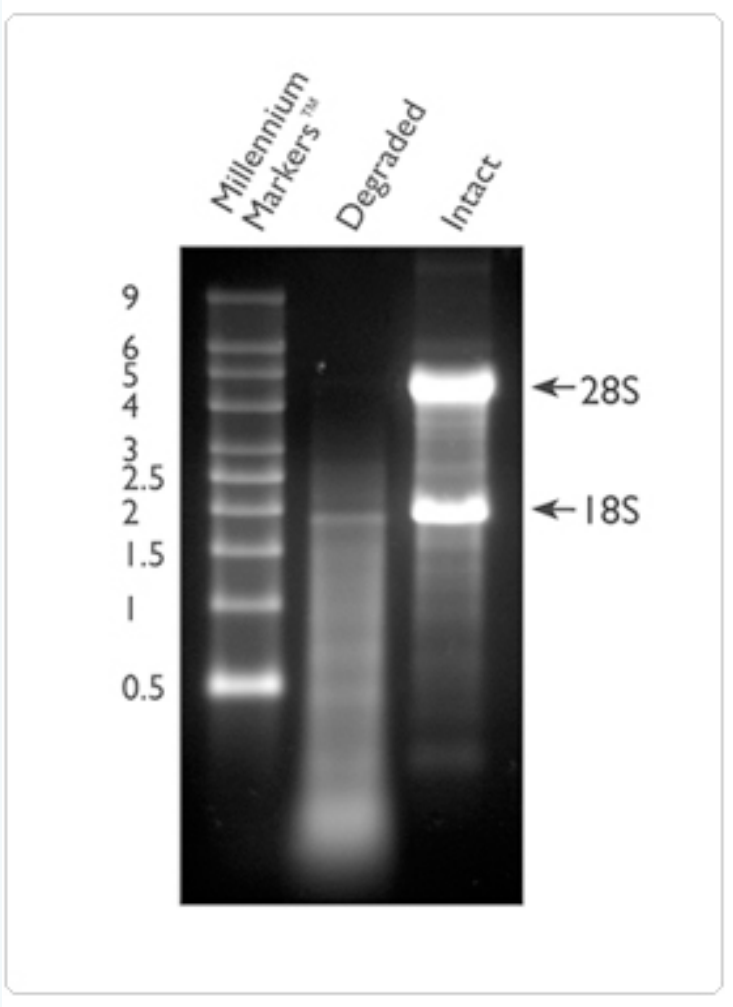
4) Use a DNase treatment step. (Residual DNA can be turned into libraries)

5) Subsample RNA after extraction so that only a small aliquot must be thawed for QC purposes. Store the main tube at -80°C until ready to begin library preparations.

RNALater vs flash-frozen tissue

- RNALater sample handling can be less rushed during RNA extraction. As soon as flash-frozen tissue thaws, RNases kick in and start degrading RNA, but RNALater has penetrated the tissue and will disable nucleases when thawed.
  - Note: this may require some testing to get right and ensure RNALater permeation. For example, some harder tissues may need to be chopped up.
- RNALater is fieldwork-friendly for when LN2 is not available. In some cases samples have gone weeks without refrigeration in warm countries and still produced decent, usable RNA.
  - That said, freezing is still the best practice for storage after RNA Later has permeated cells overnight
- Published "nucleic acid preservation buffer" (doi: 10.1111/1755-0998.12108) *very similar* to RNALater patent
- This is not to discourage flash freezing of tissue if that is your protocol but rather to encourage people doing field work to consider tissue collection methods that preserve RNA. It doesn't have to be expensive (see above)

RNA Assessment: agarose gel



RNA bioanalyzer: quality assessment RNA Integrity Number

- Algorithm designed by Agilent (Bioanalyzer, TapeStation) for their instruments.
  - https://www.agilent.com/cs/library/applications/5989-1165EN.pdf
- Based on:
  - rRNA peaks (18S, 28S) are high, indicating that rRNA is still intact
  - How much material is found in the region between the 5s and 18S regions
  - Derived from human/mouse/rat data
- Insect: https://www.agilent.com/cs/library/applications/5991-7903EN.pdf
- Plant: https://www.agilent.com/cs/library/applications/5990-8850EN.pdf
- EGL transcriptome database

RNA bioanalyzer: quality assessment
RNA Integrity Number

What the RIN can do:

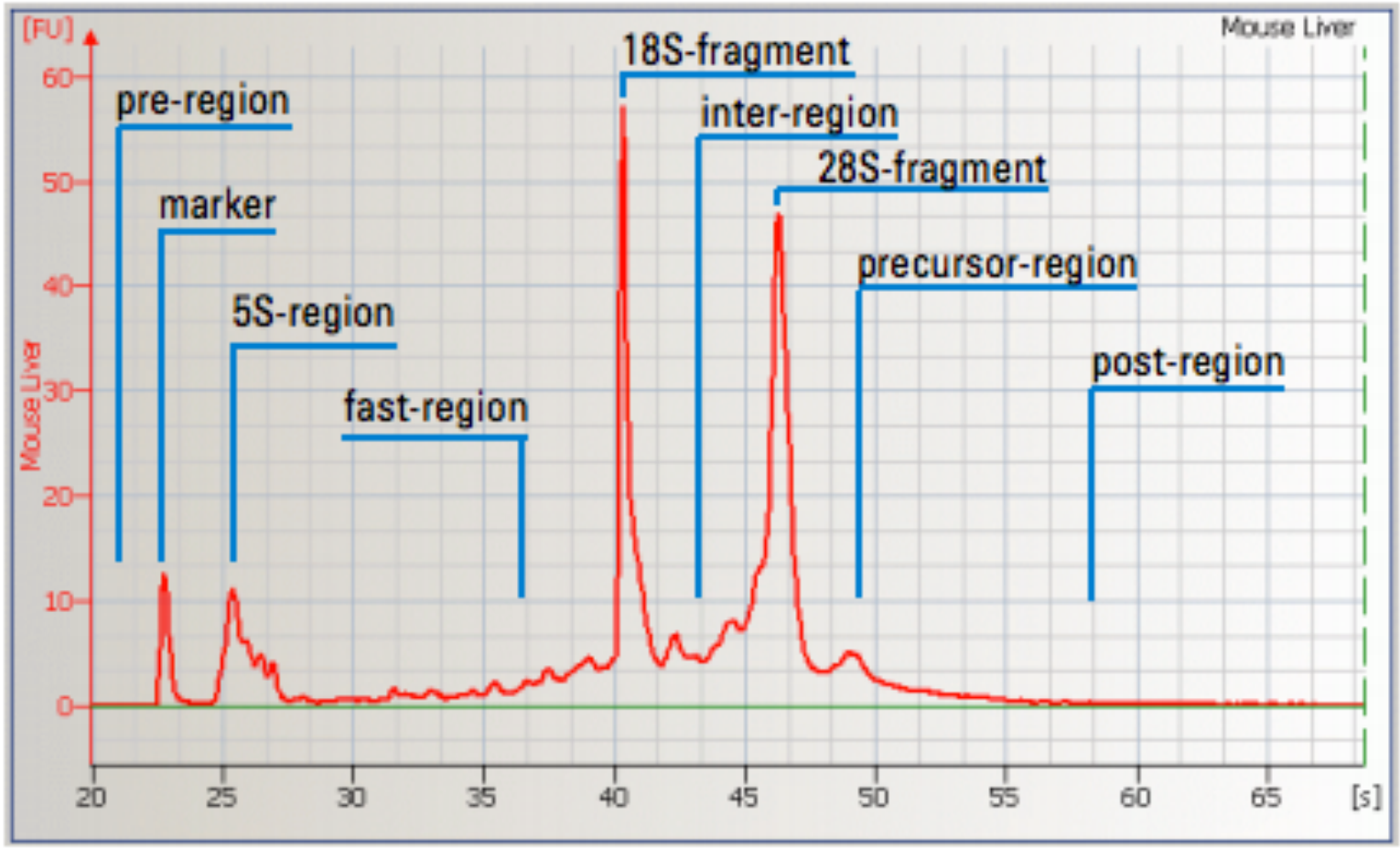
- Obtain an assessment of the integrity of RNA.
- Directly compare RNA samples (e.g. before and after shipment, compare integrity of same tissue across different labs, etc.).
- Ensure repeatability of experiments (e.g. if RIN shows a given value and is suitable for microarray experiments, then the RIN of the same value can *always* be used for microarray experiments given that the same organism/tissue/extraction method was used).

What it CANNOT do:

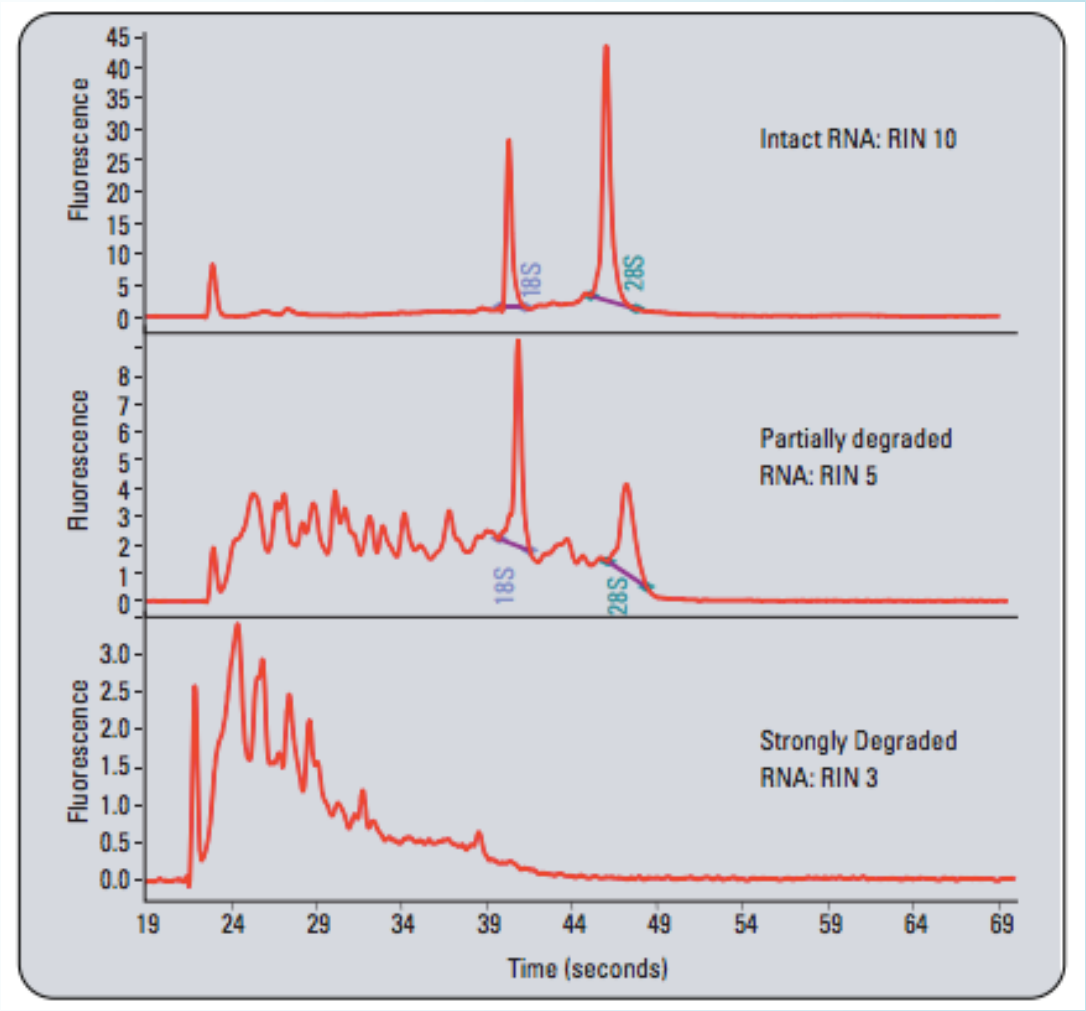
- Tell a scientist ahead of time whether an experiment will work or not if no prior validation was done (e.g. RIN of 5 might not work for microarray experiments, but might work well for an appropriate RT-PCR experiment. Also, an RIN that might be good for a 3' amplification might not work for a 5' amplification).

RIN is a good quality assessment short-cut, but it is even better to understand what the traces are telling you. In many cases, your subjective judgment is more valuable than what a single number says.

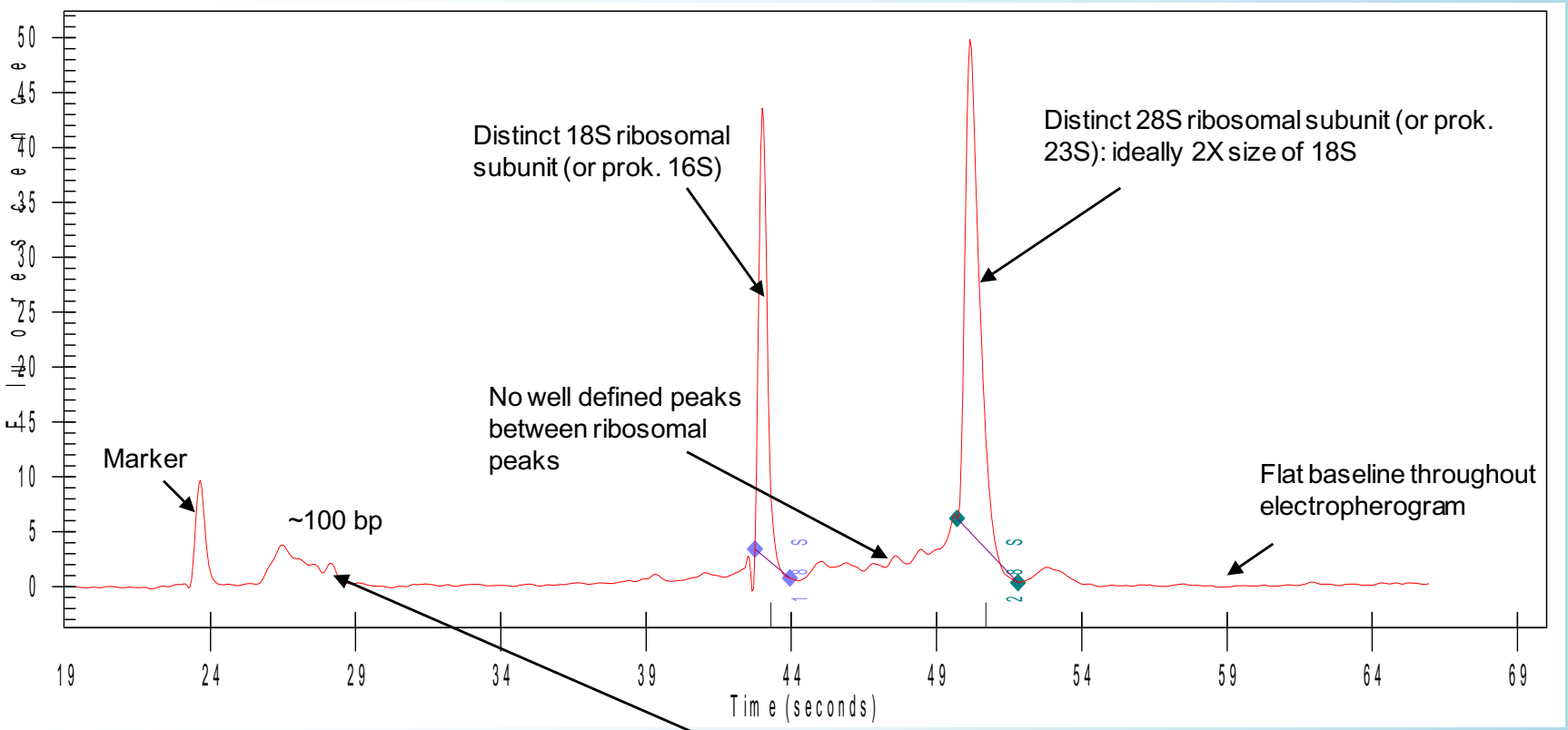
RNA bioanalyzer: quality assessment
RNA Integrity Number



RNA bioanalyzer: quality assessment RNA Integrity Number



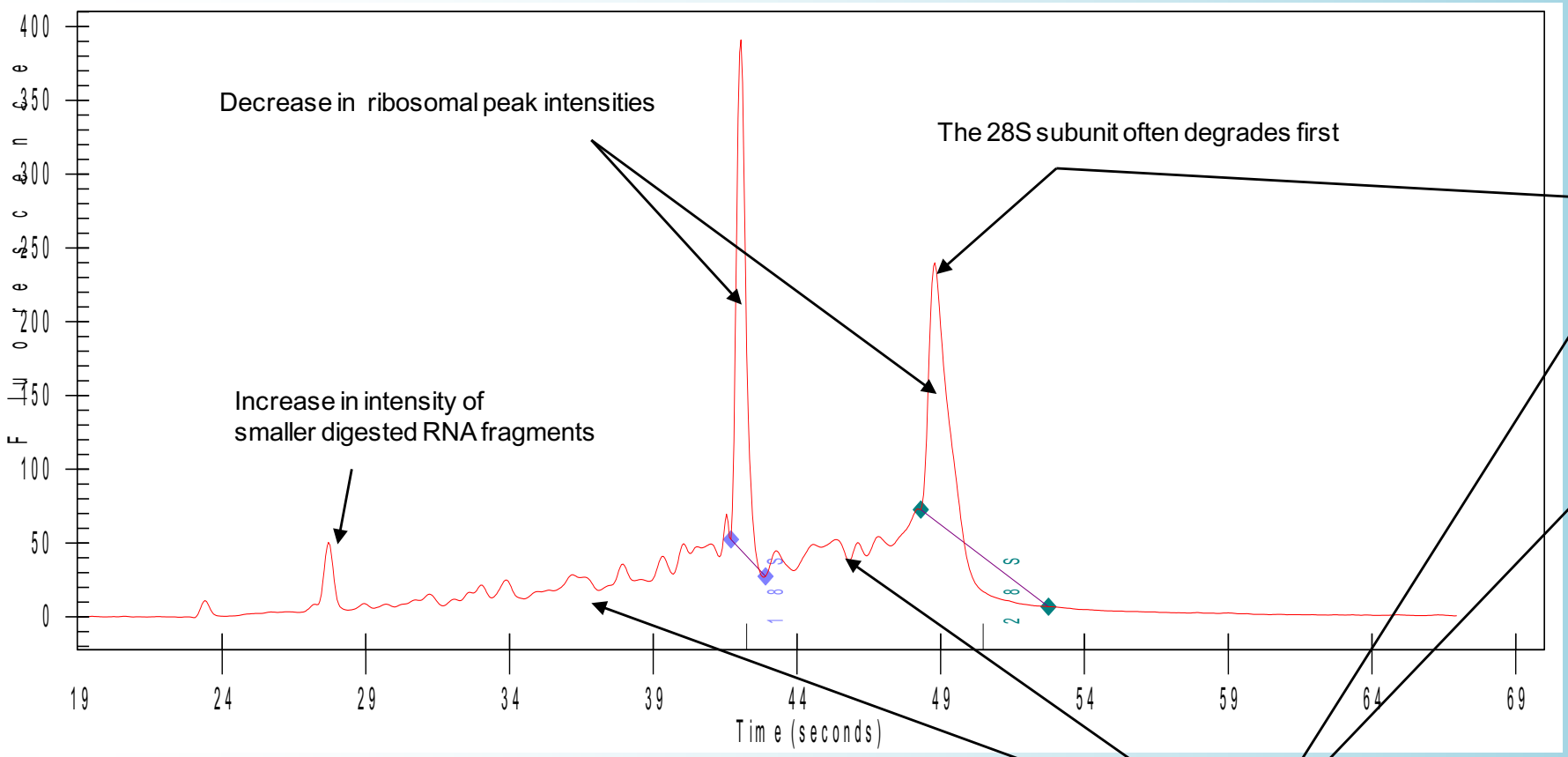
Intact Total RNA



Small peaks are sometimes present after the marker at 24 – 29 seconds. These are represented by 5S and 5.8S subunits, tRNAs, and small RNA fragments about 100bp. These are especially noted when using phenol and trizol extraction methods. They can be removed by treating total RNA through Qiagen columns which removes small RNAs.

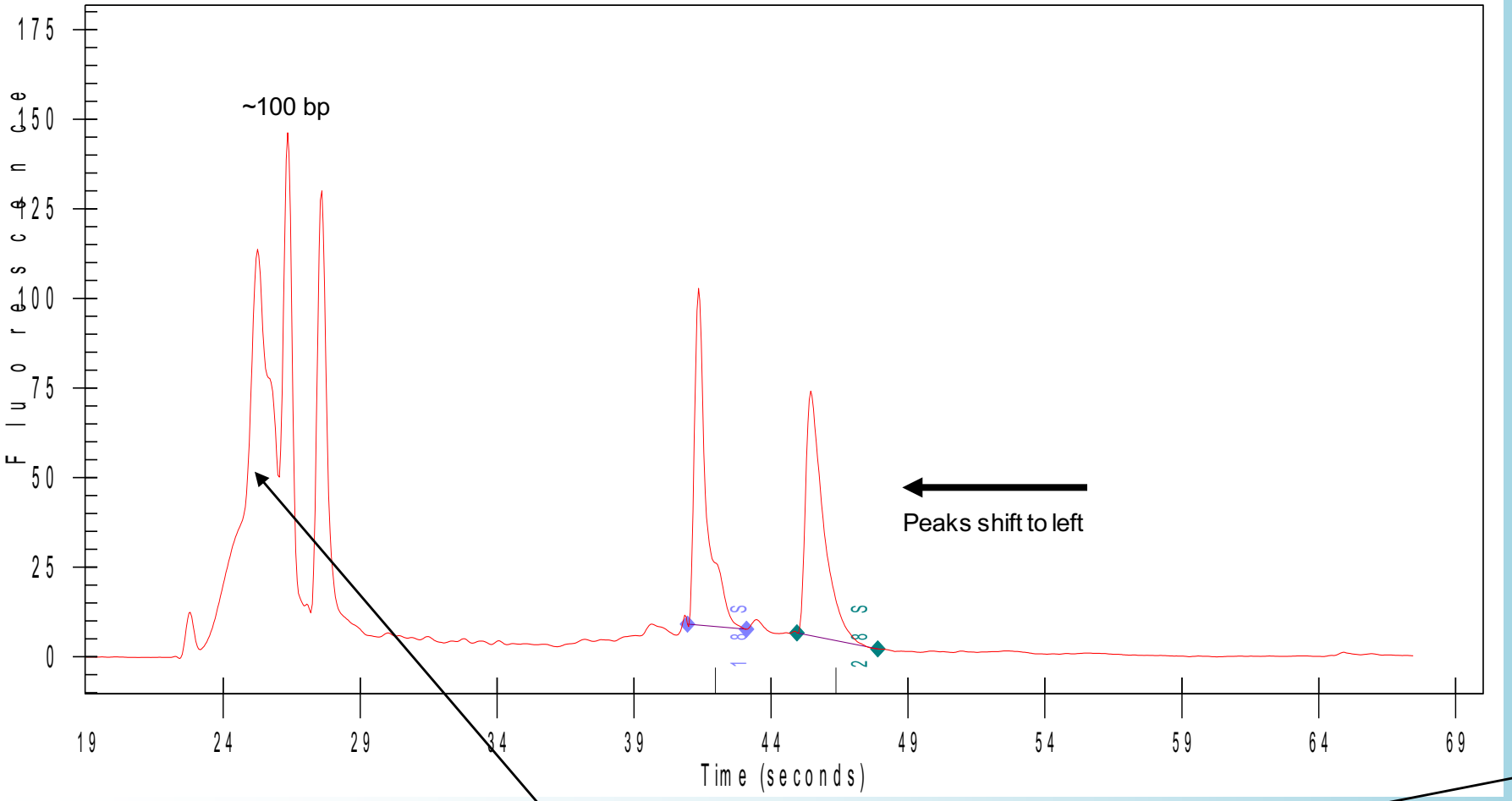
Partially Digested Total RNA

Total RNA with images like this are borderline. Re-extraction should be seriously considered.



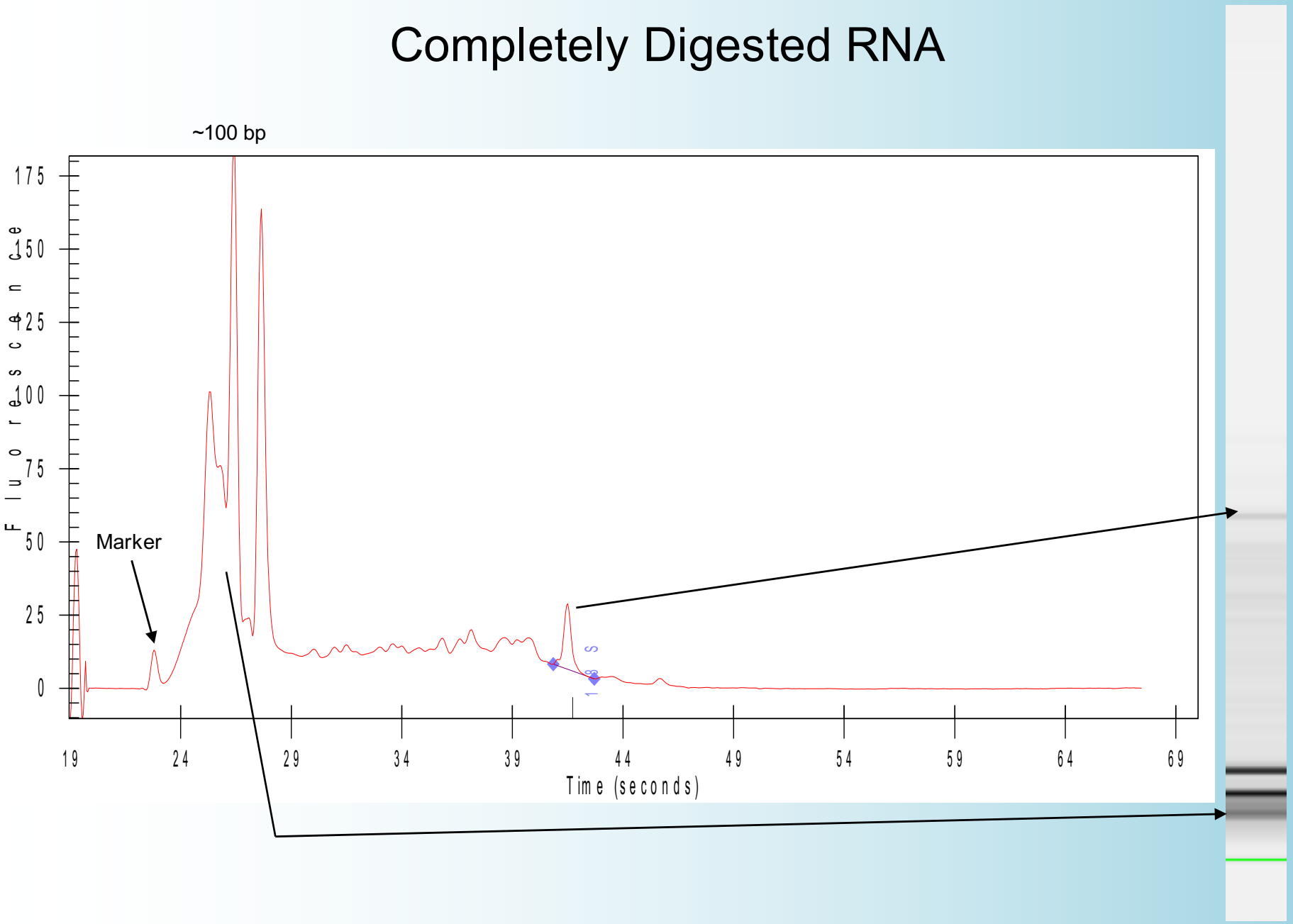
Heavily Digested RNA

Samples of this quality are questionable for use: if rRNA is digested, Poly-A tails are likely to be as well



High digested RNA peaks

Completely Digested RNA



What if my RIN is poor and/or my RNA samples appear degraded?

- Substitute other samples, if at all possible
- Determine risk of proceeding with poly-A selection:
  - looking for rare transcripts or only the most common?
  - using transcriptome data only to sequence exome or for differential gene expression?

If risk is high or if you want to sequence all RNA (viruses, small RNA, non-coding RNA in addition to mRNA), consider a ribozomal RNA depletion method

OK, I have my high quality DNA/RNA extracted and assessed in pain-staking detail, now can I start adding adapters?

Yes, but only if you are sure of your:

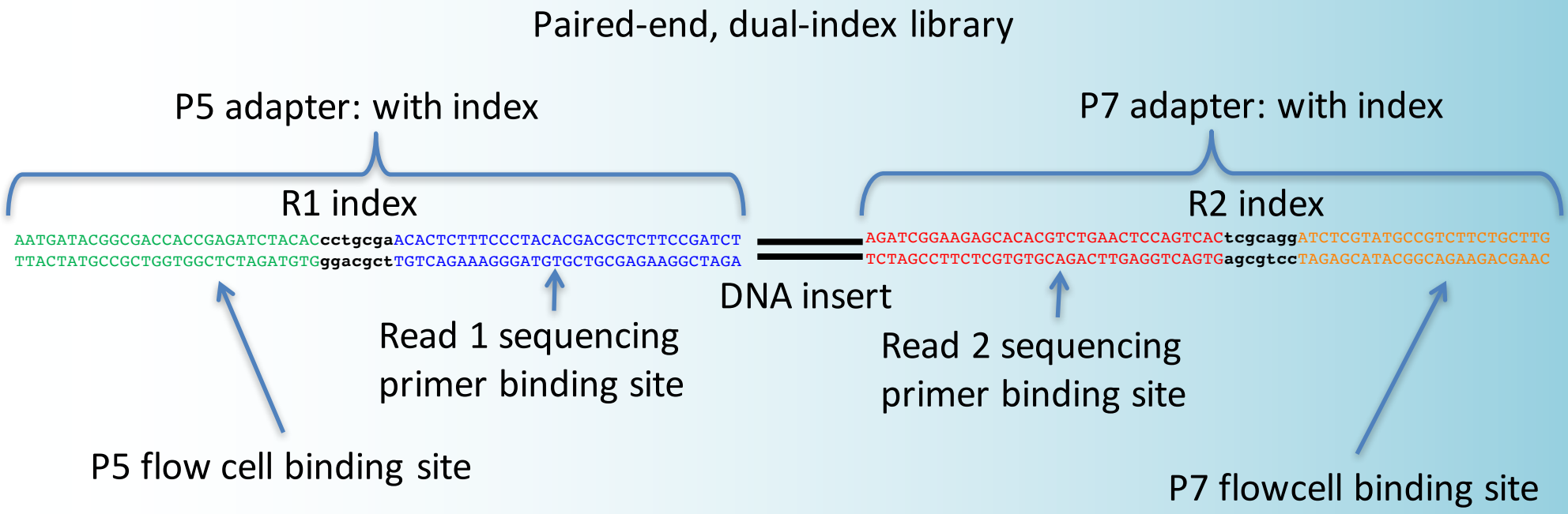
1) sequencing platform & read length
2) library preparation method
3) number of samples
4) indexing plan
5) desired insert sizing

If these factors change once libraries are begun, molecular work may have to be re-started from scratch

What is an adapter? (Illumina)

An adapter molecule has three parts:

1) Flow-cell binding sites (outer adapter)
2) Index(es)
3) Sequencing primer binding sites (inner adapter)



What does the index do?

- In Illumina sequencing, an index is a short (usually 6-8bp) unique tag of DNA incorporated into the adapter(s)
- The index sequence is read separately from the library insert, but the tag information is linked in the metadata of each sequencing read
- After a run containing multiple libraries, the data is demultiplexed: all sequencing reads with the same index tag are put into the same folder
- This process allows more than one library to be sequenced in a single Illumina lane: 2, 60, 100's …

Dual Index: Unique vs. Combinatorial

Always dual index! It costs only slightly more money for the library prep but can save a lot of money, time and hassle during sequencing.

Each sample in a project should be able to be pooled with all the other samples and be bioinformatically distinguishable from them.

Ideal: each library has a unique index on the P7 adapter and a unique index on the P5 adapter. Allows for verification of the presence/absence of chimeric molecules due to index swapping in PCR or on Illumina instruments

In the EGL, I instruct unique dual indexing for all projects where either of these two factors applies:

1) Low coverage sequencing (< 5x) is planned and samples will be run on the HiSeq 4000

2) Libraries will be sequenced immediately after adapter ligation or indexing PCR. This includes but is not limited to whole genome sequencing (most library preparation methods), RAD-seq, and RNA-seq.

Slightly less ideal: each library should *at the very least* have a unique *combination* of indexes on the P7 and P5 adapter. This approach is usually sufficient for projects where samples will undergo many processing steps after adapter ligation or indexing PCR, such as genomic DNA captures. In these cases unique dual indexing is still the best practice when possible. But it may not always be possbile due to samples sizes and adapter availability.

With either approach, be meticulous about your recordkeeping. The index is the only link between your sequencing data and its biological meaning

Index incorporation

Indexes are incorporated into the library adapter in two ways as you will see in the following section:

1) Two-step, Stub + extension
2) One-step, Full-length

For labs that aren't high-volume and don't need a lot of each index, it isn't usually cost-effective to purchase them individually.

Instead, some library prep kits come with indexes (often 12 single or 96 combinatorial dual) and indexes can often be purchased from vendors or larger labs that have a greater volume of library preps (some options are noted in the following slides)

Selecting indexes

1) Each library in a project should have a unique index (better yet, **two** unique indexes: one on the P5 adapter and the other on the P7) so that they can all be combined and run in the same sequencing lane(s)

2) If planning a capture experiment, the indexes you use must be compatible with available blocking oligos

3) Be careful to keep meticulous records with respect to which index gets incorporated into which library.

Balancing indexes

Index combinations should:

1. Have sequences as distinct from each other as possible
2. Have a relatively even representation of all 4 bases at each position
3. Must have A/C and G/T in each position of the index

Table 4: Examples of Good and Bad Index Combinations

Good Examples				Bad Examples			
Index 1		Index 2		Index 1		Index 2	
705	GGACTCCT	503	TATCCTCT	705	GGACTCCT	502	CTCTCTAT
706	TAGGCATG	503	TATCCTCT	706	TAGGCATG	502	CTCTCTAT
701	TAAGGCGA	504	AGAGTAGA	701	TAAGGCGA	503	TATCCTCT
702	CGTACTAG	504	AGAGTAGA	702	CGTACTAG	503	TATCCTCT
	✓✓✓✓✓✓✓✓		✓✓✓✓✓✓✓✓		✓✓✓✓✓✓✓✓		✓✓✓✓XXXX

✓= signal in both color

X= signal missing in one color channel

Balancing indexes

This is especially important when using just a few (2-8) indexes at either the P7 or P5 position because the odds are higher that by chance you will pick a poor combination

You can usually find low-plex pooling information from the company that provides the indexes or the protocol that they were synthesized from.

Or check with the published protocol or lab you purchased indexes from. For Meyer & Kircher indexing oligos used in the EGL, if used in order starting with indexing1, they will be balanced

External Index vs Internal Barcode

- External index is read separately from sequencing data and is bioinformatically linked
- Internal barcode is read by the sequencer at the start of the read.
- A small amount of read length is lost with an internal barcode, but that is off-set in some projects by these advantages:
  - Combinatorial barcoding allows unique combinations of internal barcode + external adapter. (Ex: if you have only 24 internal indexes and 10 external indexes, 240 samples can be pooled in the same sequencing lane.)
  - In two-step adapter incorporation methods, samples can be pooled earlier in the library prep process to save money
  - Internal barcode creates sequence diversity at crucial initial bases for RAD-Seq and amplicon libraries (most other types of libraries have naturally high complexity)

How to incorporate the adapter and index onto DNA?

Depends on the type of data to be collected:

- genomic DNA (whole genome sequencing, targeted capture)
- RNA sequencing
- RAD-Seq
- Amplicon sequencing
- ChIP-Seq
- Mate Pair

Genomic DNA library preps

The biggest time investment in molecular lab work is not in adding adapters. It is in the extraction, assessment and preparation of nucleic acid to begin.

- DNA extraction
- Agarose gel electrophoresis
- Qubit assessment
- Dilutions
- Sonication (Bioruptor or Covaris) and assessment (gel electrophoresis or Bioanalyzer)*
  - Repeat these steps as needed
- Bead clean-up and size-selection*

*Note: these stages are skipped when using Nextera kits

This is where the labor goes in to get things right (it can't be corrected later)

Time spent obtaining high-quality DNA, fragmented to the correct size will pay off

Enzymatic steps of library protocols are relatively rote and easy once the DNA is ready

DNA fragmentation: gDNA
Two-stage library preparation & Y-shaped adapter

Sonication:

- Covaris is available in the FGL
  - More reliable with proper sizing but consumables are expensive ($5-10/tube)
- Bioanalyzer is available in the EGL
  - Consumables are cheap ($0.30) but this instrument can be frustratingly inconsistent

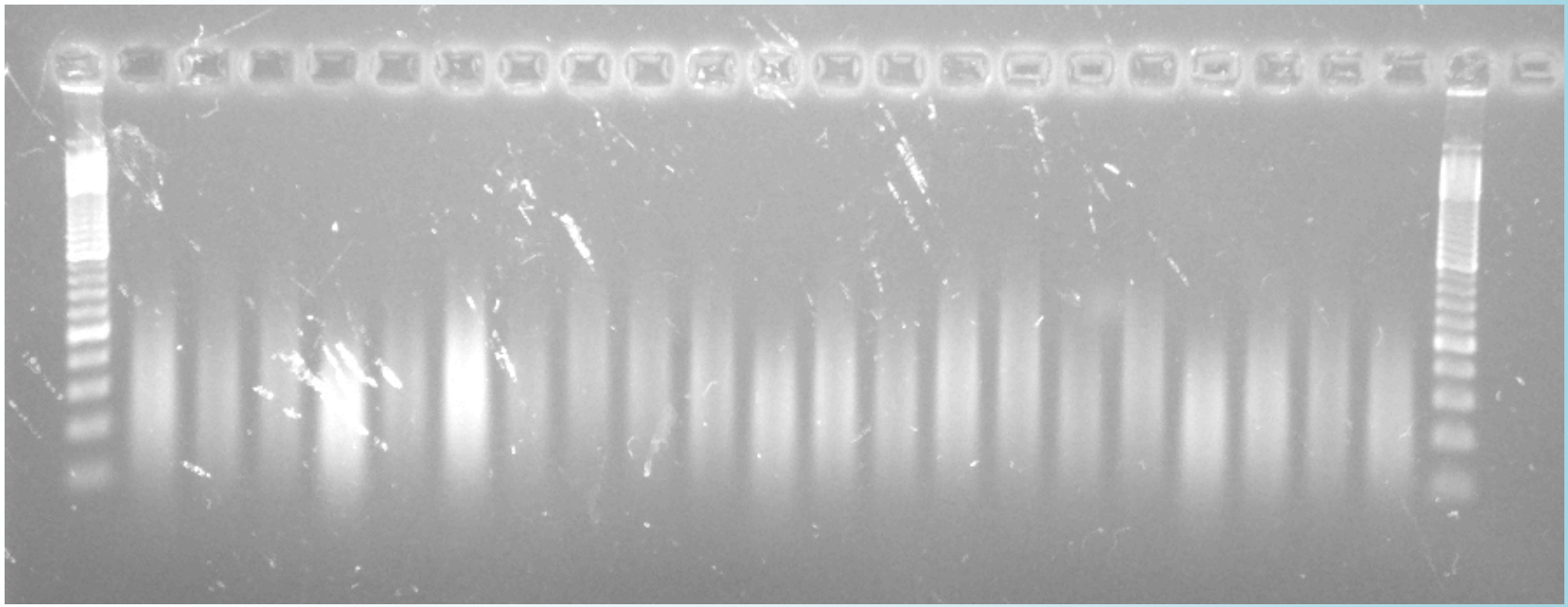
Fragmentase:

- Sold by New England Biolabs
- Part of Kapa Biosystems HyperPlus Kit

Both enzymatic and sonication methods may require a fair amount of optimization to figure out the right conditions. However, if all samples are handled the same way (concentration, dilution buffer, etc), hopefully those conditions, once determined, can be reliably used for most samples of a project.

DNA fragmentation: gDNA

The ideal DNA smear distribution will depend on the type of data being collected, the length of the sequencing run planned, and whether you want PE reads to overlap or prefer that they don't

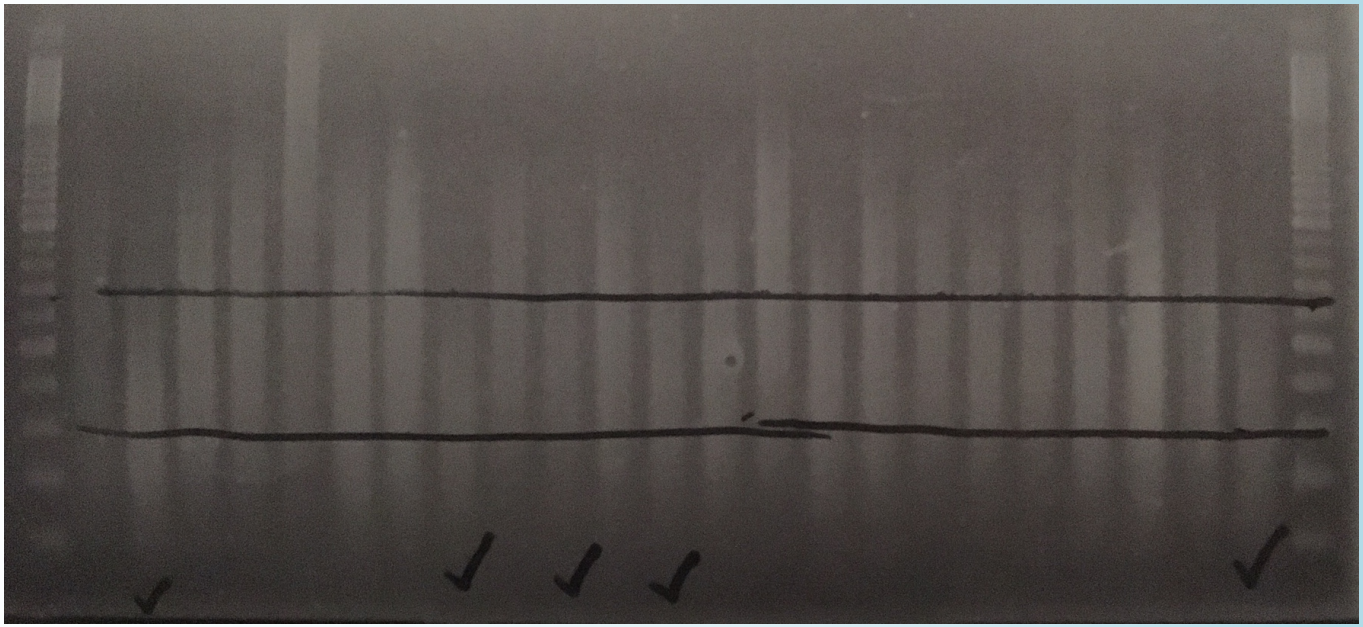


Final post-Bioruptor sonication gel image for planned exon capture experiment to be sequenced with PE100. In this case the aim is for an average insert size around 250-300bp, large enough so that most reads do not overlap in the middle but small enough so that insert sizes stay relatively small (smaller libraries tend to capture better)

DNA fragmentation: gDNA

The Bioruptor can be frustrating and tedious since we must err on the side of under-shearing to start.

Covaris (FGL) has a reputation as more consistent but consumable costs add up with larger projects



In progress post-Bioruptor sonication gel image for WGS libraries to be sequenced with PE150. In this case the aim is for an average insert size around 300-600bp so that most reads do not overlap in the middle. After 5 cycles of sonication (30s LOW/90s OFF) only 5 of 23 samples are at the proper size. All others repeated for 2 cycles—gel will be re-run…

Illumina library preparation: gDNA
1) Two-stage library preparation

--Meyer & Kircher 2010: doi:10.1101/pdb.prot5448

--iTru/Adapterama system:
https://conf.abrf.org/sites/default/files/images/tglenn_abrf_adapterama_march_2015_final.pdf

--NEBNext

Pros:

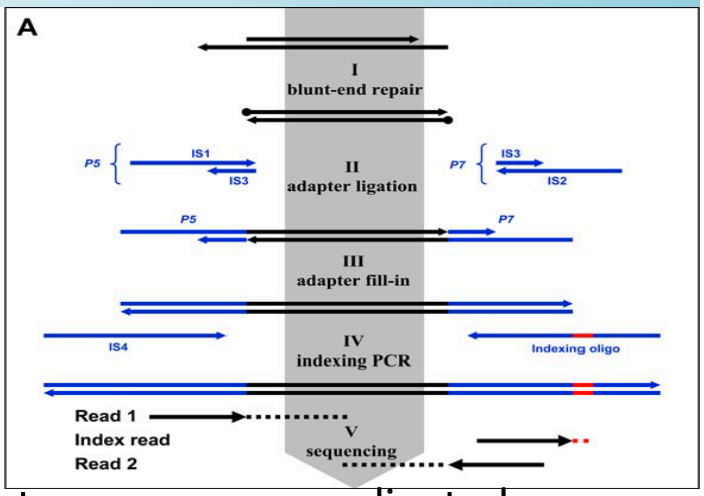
- No kit required (reagent costs in EGL = $10-15 per sample) or allows the use of less expensive kits
- Very inexpensive: internal stub adapters are non-indexed: index is incorporated through PCR with longer indexing oligos
- Flexible design makes it easy to add additional indexes to a lab's database or to incorporate dual indexing
- Multiplexing: the sky's the limit! Combinatorial dual indexing allows 100s of samples to be pooled together on a single lane. These are expensive to buy in large quantities but can be purchased from other lab groups:
  - EGL (Meyer & Kircher protocol): 96 x P7 indexes; 24 x P5
  - GSL: sells adapter stubs and plates of indexing oligos
  - Glenn (iTru): 387 P7; 192 P5 (http://baddna.uga.edu/)
- Degraded DNA/historical samples work just fine (this method often used in aDNA labs)

Cons:

- Best practice is 1000 ng of starting material if libraries will be captured (usually easy to come by but not in all cases).
- Less efficient at converting fragments into final libraries than the methods that will follow
- Has more steps than the other gDNA methods, so more time in the lab
- Unique dual indexing of all samples cost-prohibitive
- Requires use of a sonicator or fragmentase
- PCR step is required to extend the adapter and incorporate the index (only a con if a no-amplification method is necessary)

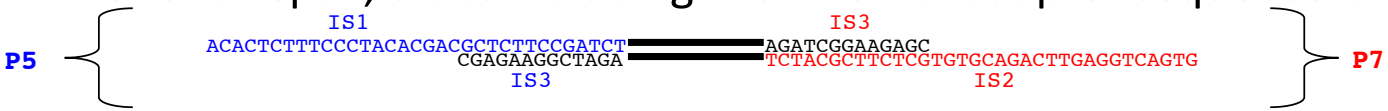
Illumina library preparation: gDNA
1) Two-stage library preparation

Meyer & Kircher protocol (2010)



Meyer, Matthias, and Martin Kircher. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing." *Cold Spring Harbor Protocols* 2010.6 (2010) DOI: 10.1101/pdb.prot5448

Adapter ligation: 1) After end repair, stubs including the internal adapter sequence are ligated

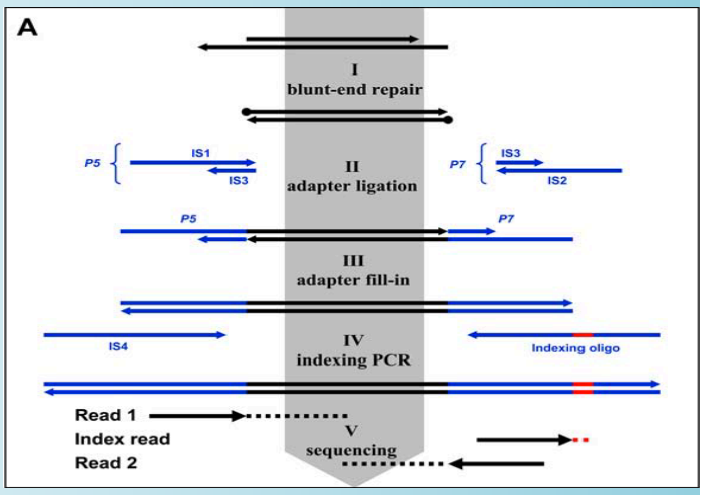


Adapter fill-in: 2) Internal adapters are filled in to be double-stranded

Indexing PCR: 3) Adapters are extended to full-length and indexes are incorporated with PCR

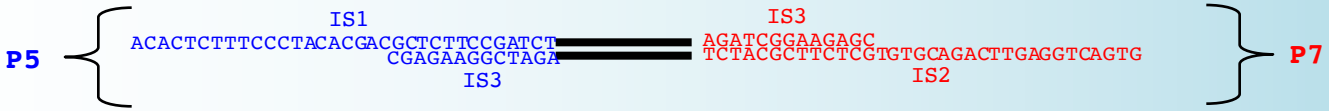
Library with adapters:

Kircher, Sawyer & Meyer (2012) dual index protocol

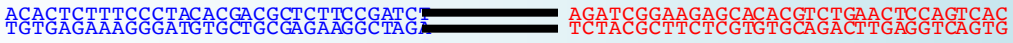


Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40: e3. doi: 10.1093/nar/gkr771.

Adapter ligation:

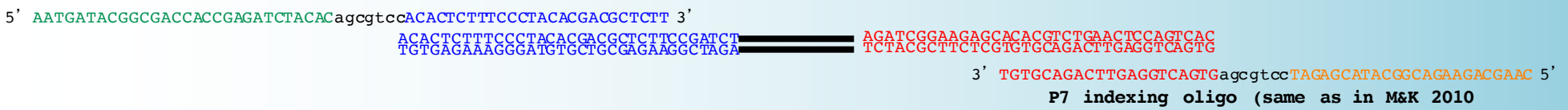


Adapter fill-in:



Indexing PCR:

P5 indexing oligo (new: Kircher, Sawyer & Meyer 2012)



P7 indexing oligo (same as in M&K 2010

Library with adapters:

Same thing but with two indexed adapters

Illumina library preparation: gDNA
2) Y-shaped adapter

--Illumina TruSeq kit (comes with 12 single-index adapters, Illumina sells unique dual index adapters in sets of 24 or 96)

--Kapa Biosystems HyperPrep/HyperPlus (adapters not included but may be purchased in sets of single index 12 & 24, or combinatorial dual index 96)

--many other vendors: NuGen, NEB, etc.

Pros:

- Slightly fewer steps than the two-stage library prep; higher yields
  - Kapa Hyper kits reduce the number of steps even further.
- PCR enrichment step is not necessary since the library is full length after adapter ligation; can be used for library preps when PCR should be avoided or minimized so as not to introduce biases
- Best choice for WGS, especially with large genomes. An amplification-free approach eliminates PCR biases caused by difficult to amplify genomic regions
- Degraded DNA, historical samples work just fine
- Samples can be used for downstream enrichment by targeted capture

Cons:

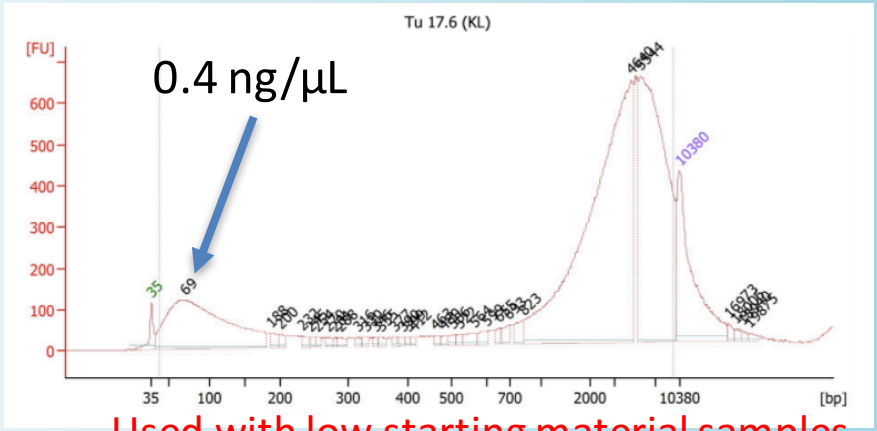
- PCR-free requires around 200 ng of starting material after sonication and size-selection (usually easy to come by but not in all cases)
- Adapters are expensive since each one must be individually barcoded. (From GSL cost is about $5/library for 5 µL @ 15 µM)
- Using kits is more expensive per library than a homebrew method
- Requires use of a sonicator or fragmentase

Illumina library preparation: gDNA
2) Y-shaped adapter

- Higher yield of conversion of unique fragments into full libraries
  - https://doi.org/10.1186/s12864-016-2757-4
- PCR step is optional and can be skipped for WGS to avoid amplification biases
- Kits allow lower throughput labs to not have such a large start-up investment
- Fewer steps increase yields and save researcher time

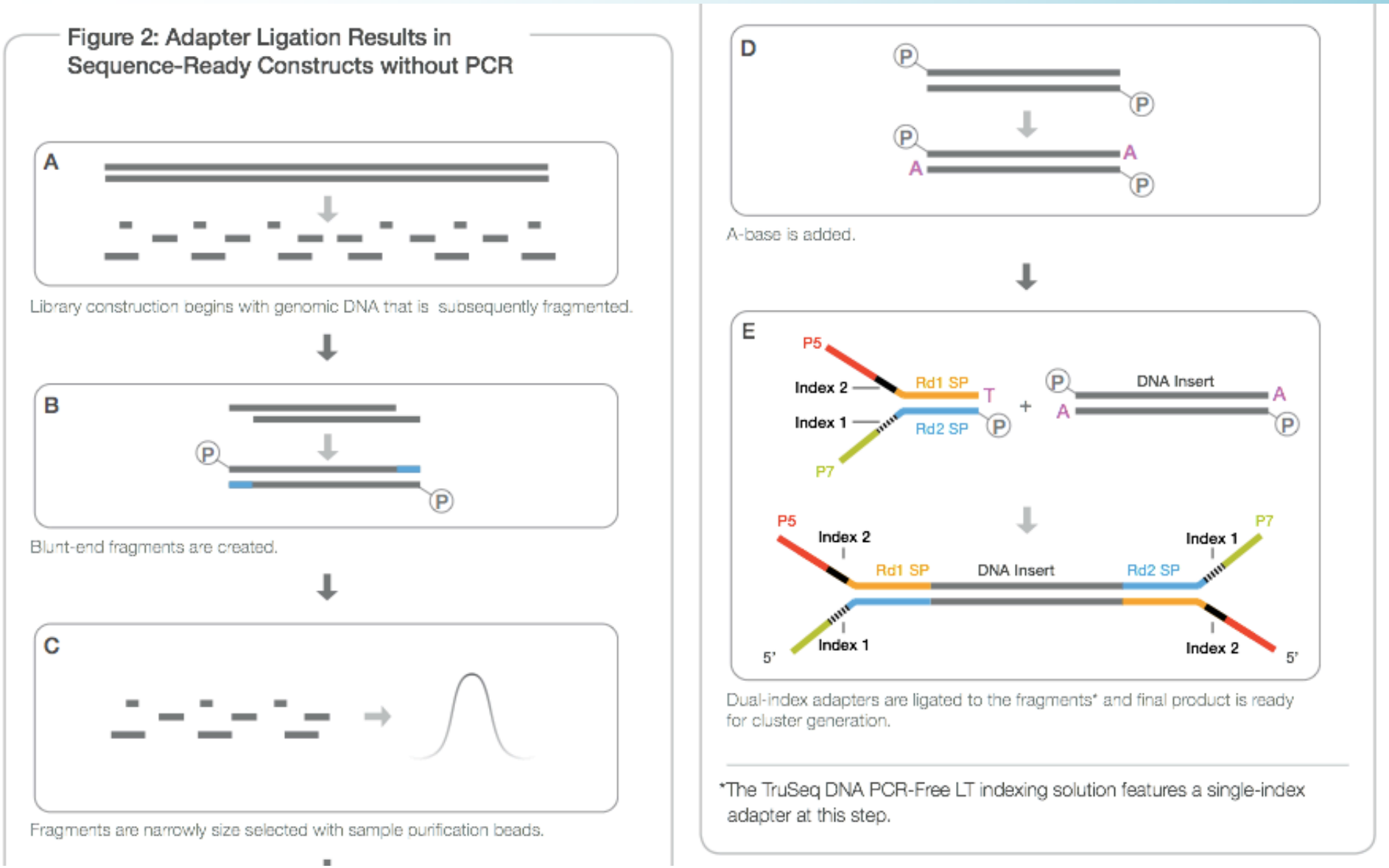
Hands-on time	Total time	
5 min	5 min	Reaction Setup
0 min	60 min	End repair and A-tailing
0 min	15 min	Adapter ligation
15 min	30 min	SPRI cleanup
0 min	30 min	Library Amplification
15 min	30 min	SPRI cleanup

Kapa Hyper Prep:
Only 3 enzymatic steps + 2 bead clean-ups after sonication



Used with low starting material samples

Illumina library preparation: gDNA
2) Y-shaped adapter



Illumina library preparation: gDNA
3) Nextera

Illumina Nextera: reagent and index costs ~$100/sample (smallest kit: 24 reactions)

--Preps are very expensive but enzymes can be radically reduced to stretch them (5-10 fold!)

--Problem: DNA input requirement is already low (50ng) and must be proportionately reduced when diluting the kit (cut frequency based on concentration of DNA).

Pros:

- Least number of hands on steps in the lab→ very fast (as low as 90 minutes of lab time)
- Does not require a sonicator or fragmentase
- Protocol requires very small amounts of starting material. Good news for smaller organisms that have smaller genome sizes (drosophila!)

Cons:

- Cannot be used with degraded/historical DNA
- Very low starting material (50ng with full protocol, as low as 5ng for modified), so there are fewer copies of the genome turned into libraries.
- Not recommended for downstream targeted capture due to low starting material
- PCR step is required (only a con if a no-amplification method is desired)
- May display some biases in transposase cut sites
- Most expensive method, only sold by Illumina.

Illumina library preparation: gDNA
3) Nextera

Transposase:

1) Fragment
2) End Repair
3) Ligate Internal Adapters

All in one enzymatic step!

Then the index and external adapter is incorporated with PCR

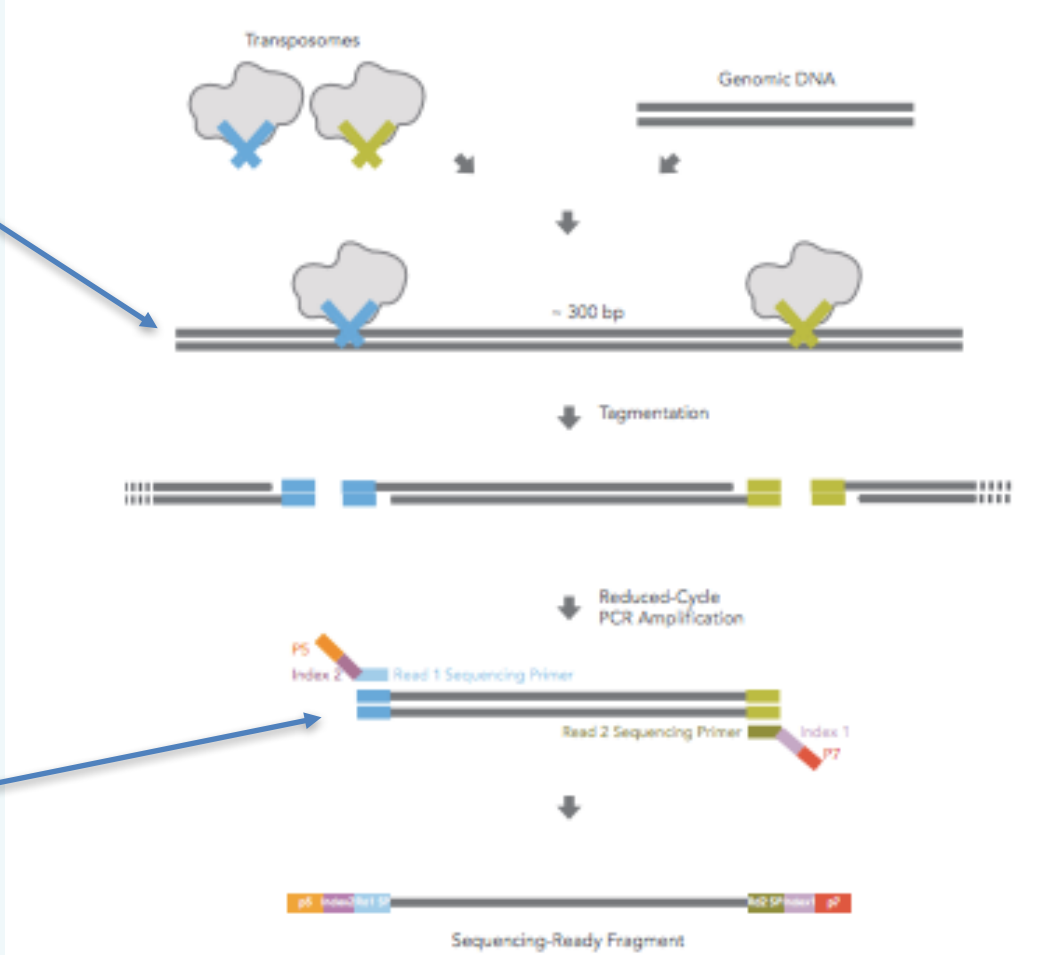


Figure 2: Nextera Library Preparation Biochemistry—Nextera chemistry simultaneously fragments and tags DNA in a single step. A simple PCR amplification then appends sequencing adapters and sample indexes to each fragment.

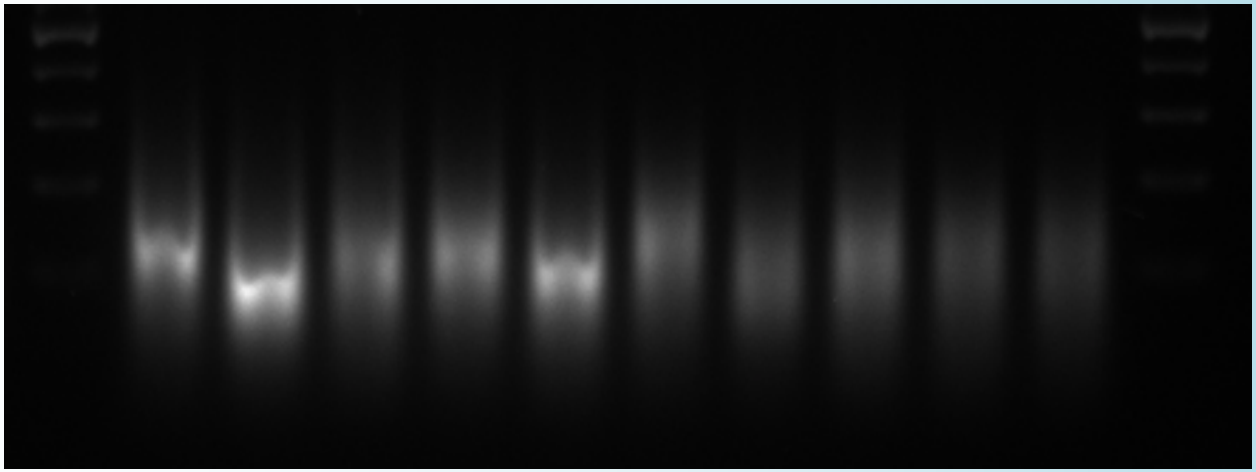
Researcher Advice: Nucleic Acid/Library Prep

- *If seeking tissues for HMW DNA extractions, save tissues other than liver (which is full of nucleases that can shorten DNA fragments)*
- *For users of the Bioruptor: budget more time than you think for sonification/run gel/re-sonification/re-run gel process.*
- *Take the time to get a good handle on what each step in the library prep process accomplishes at the molecular biology level and the ways in which decisions during library prep can affect the eventual sequencing data. That would have given me more confidence to be able to troubleshoot more effectively in the lab and made the lab work less daunting and more efficient.*
- *Do a test-run of your protocol with a few DNA or RNA samples that you don't care about. It's really helpful for practice and streamlining your procedure, especially for all the steps that aren't written in protocols - like gathering all the right supplies for library prep before starting.*
- *Never take the lab work for granted. Bioinformatics isn't magic and can only rarely solve inappropriate study design or sloppy lab practice. Spending a few extra days reextracting DNA and/or redoing a library is trivial compared to the headache later along the line because you "couldn't be bothered" at the time.*
- *The Meyer and Kircher library prep protocol is very very robust. Most likely it will work. Relax!*
- *HTS is more robust than PCRs & Sanger sequencing. Also, it is so much fun.*

Museum/Historical Samples

- DNA extracted from tissue sources not collected or preserved for that purpose but for other scientific study in museums
  - study skins, hair, toe pads, formalin-fixed specimens, pinned insects/spiders, herbaria specimens
- Historical DNA is a subset of ancient DNA
  - DNA recovered from biological samples that have not been preserved specifically for later DNA analyses
- Older ancient DNA samples have been exposed to the environment longer after death and require more specialized lab facilities for safe handling (positive airflow, UV lamps, full body PPE.)
- Moderate quantities of DNA can be extracted from museum samples of the past ~100 years
  - need for an isolated space, separate reagents, and extreme caution when handling samples.

Historical/Degraded DNA

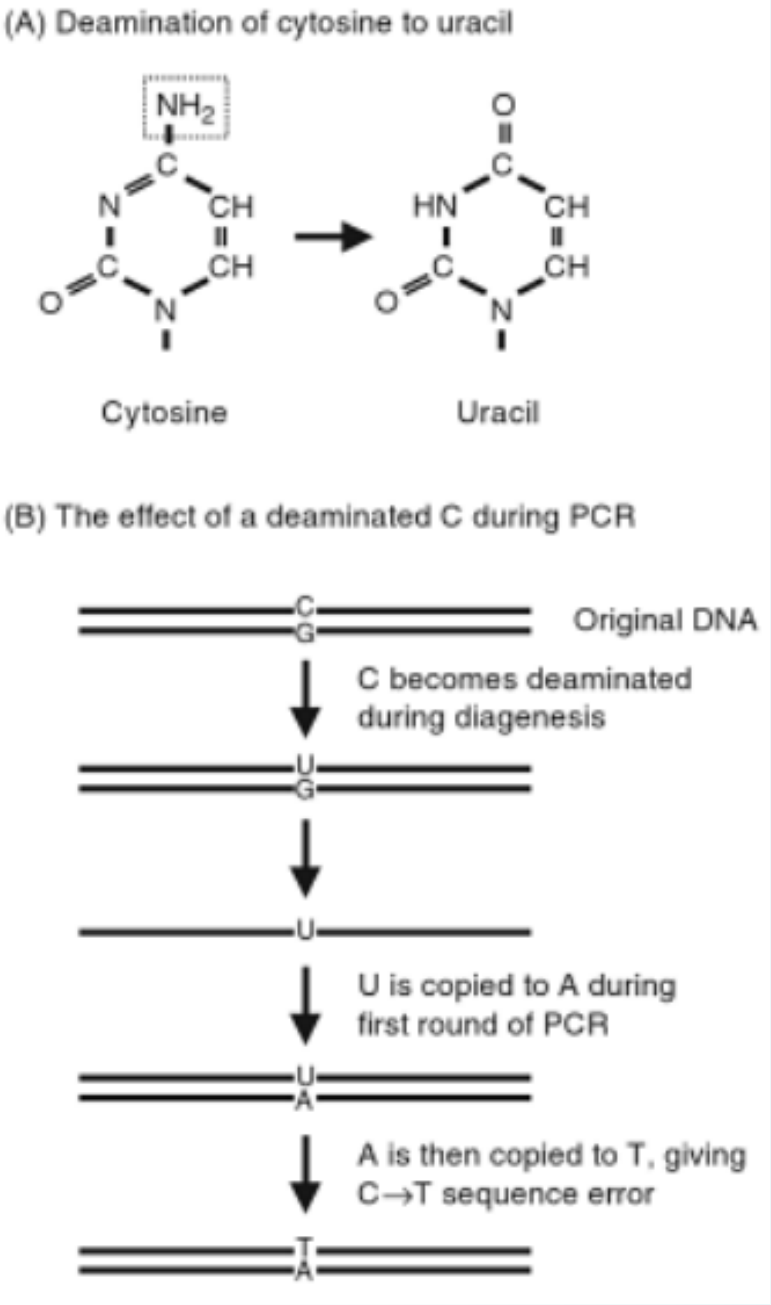


- Degraded samples may still be appropriately sized for short-read sequencing
  - Often they do not even need sonication and are much faster to prepare!
- Ideal for targeted capture experiments
  - Captures reduce the amount of contamination and enrich for endogenous material
  - Smaller inserts tend to capture even better than longer modern ones
- If more than partially degraded, not a good fit for RAD-Seq unless combined with a capture approach (HyRAD, Rapture)
- Impossible for RNA-Seq or de novo WGS (with no available reference)

Historical DNA: challenges

- Destructive sampling
  - often museums wary of "loaning" materials without proof-of-concept for the work
- Sometimes samples are **very** fragmented and produce **very** low yields. Difficult to trouble-shoot/optimize
- Higher likelihood of environmental and adapter contamination reducing the proportion of usable reads
- Tend to have much higher duplicate sequence rates and therefore need more reads to reach the same coverage level as modern samples.
  - These can be bioinformatically removed, but result in loss of data
- Some preservation methods introduce contaminants and complications beyond fragmentation of DNA
  - Ex. Formalin-fixed materials extra-challenging due to cross-linking with protein in addition to degradation. DOI: 10.1371/journal.pone.0141579

Historical DNA: challenges



Post-mortem DNA damage caused by hydrolysis often converts cytosine to uracil. Subsequent polymerases replace uracil with thymines giving the appearance of a spurious C → T substitution (or G → A in reverse strand)

1) Use proofreading polymerases that stall in the presence of uracil to reduce (but not eliminate) misincorporation errors [note: you should use proofreading polymerase for all library preparations involving PCR]
2) These substitutions can be filtered from the analysis
3) Ends of reads can be trimmed if samples were not sonicated (but that may mean a significant % data loss)

Historical DNA: USER enzyme treatment

Jeff Good lab (Univ of MT) protocol, based on:

- Shapiro B, Hofreiter M. 2012. Ancient DNA: Methods and Protocols.
- Briggs et al. 2010. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Research 38:e87. https://doi.org/10.1093/nar/gkp1163

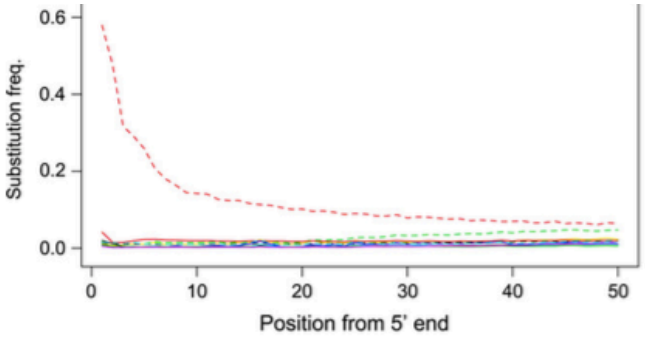
USER is an New England Biolabs enzyme mix of:

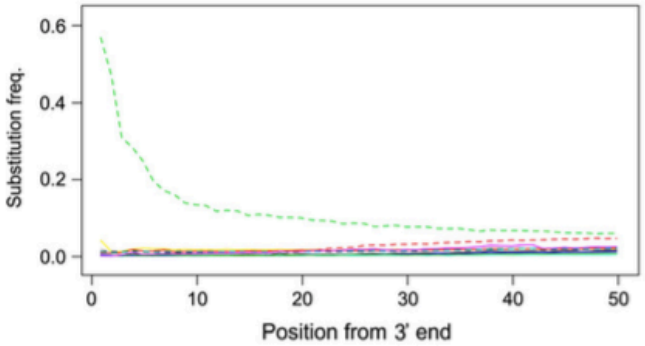
- Uracil DNA glycosylase (catalyses the excision of a uracil base)
- Endonuclease VIII (cleaves the DNA phosphodiester backbone to release the two fragments)
- Cost ~$4.00/reaction

Remainder of library prep after end repair can follow two-stage adapter incorporation (Meyer-Kircher) or Y-shaped (Kapa/TruSeq)

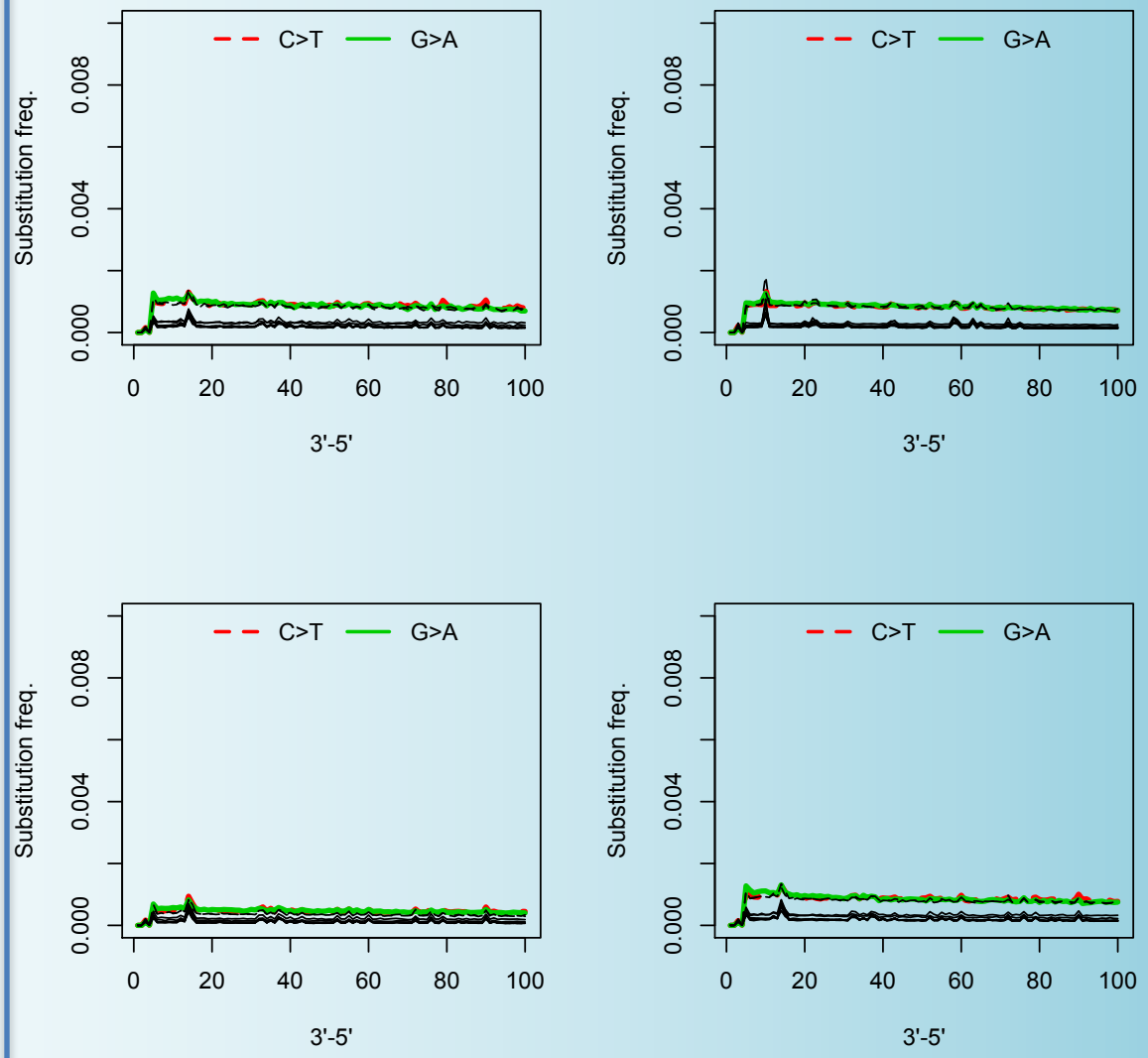
Historical DNA: USER treatment results

Chipmunk skin extractions
(Not USER-treated)





Bi, et al. 2013
Unlocking the vault:
next generation
population genomics
doi: 10.1111/mec.12516



4 squirrel skin extractions (USER-treated)

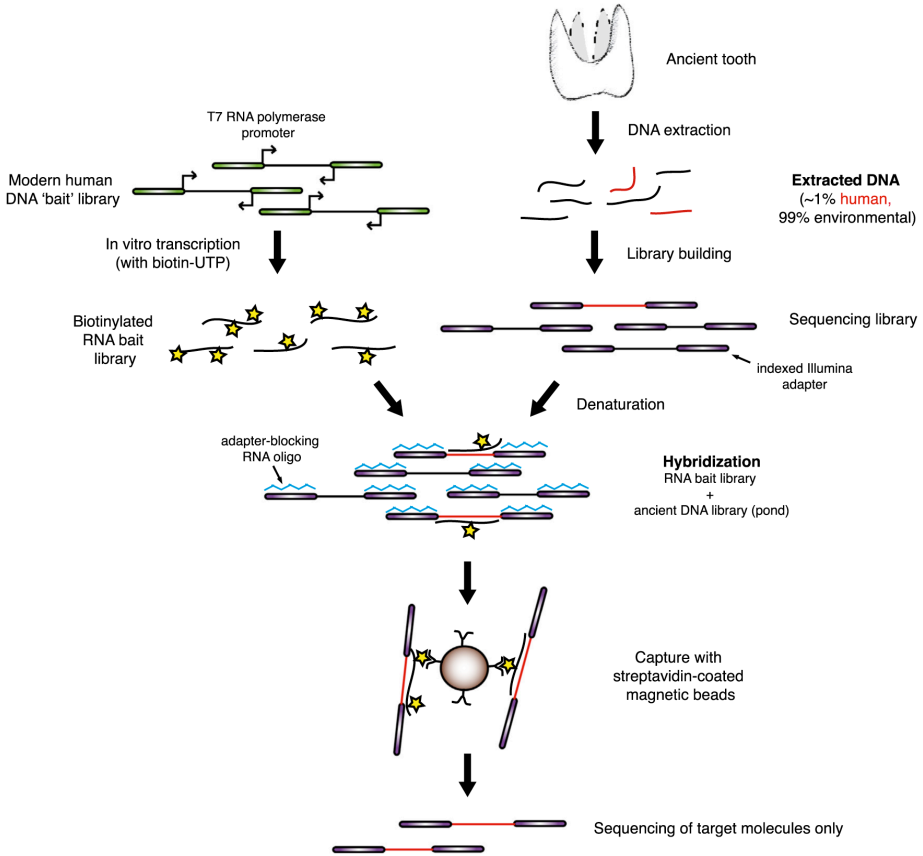
Target Enrichment - Whole Genome In-Solution Capture

Please cite this article in press as: Carpenter et al., Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries, The American Journal of Human Genetics (2013), http://dx.doi.org/10.1016/j.ajhg.2013.10.002

ARTICLE

Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries

Meredith L. Carpenter,[1] Jason D. Buenrostro,[1,14] Cristina Valdiosera,[2,3,14] Hannes Schroeder,[2] Morten E. Allentoft,[2] Martin Sikora,[1] Morten Rasmussen,[2] Simon Gravel,[4] Sonia Guillén,[5] Georgi Nekhrizov,[6] Krasimir Leshtakov,[7] Diana Dimitrova,[6] Nikola Theodossiev,[7] Davide Pettener,[8] Donata Luiselli,[8] Karla Sandoval,[1] Andrés Moreno-Estrada,[1] Yingrui Li,[9] Jun Wang,[9,10,11,12] M. Thomas P. Gilbert,[2,13] Eske Willerslev,[2,15] William J. Greenleaf,[1,15,*] and Carlos D. Bustamante[1,15,*]



WISC works effectively for enriching genomic DNA from ancient specimens that contain very low levels of endogenous DNA (<1%)

Can increase unique reads 2- to 10-fold

Probes: prepared from modern genomic DNA converted to RNA or synthesized by commercial vendors (Arbor/Mycroarray)

doi: 10.1016/j.ajhg.2013.10.002

What to do with genomic DNA libraries? Sequence the whole thing or capture a portion of the genome?

Whole Genome Sequencing (re-sequencing)

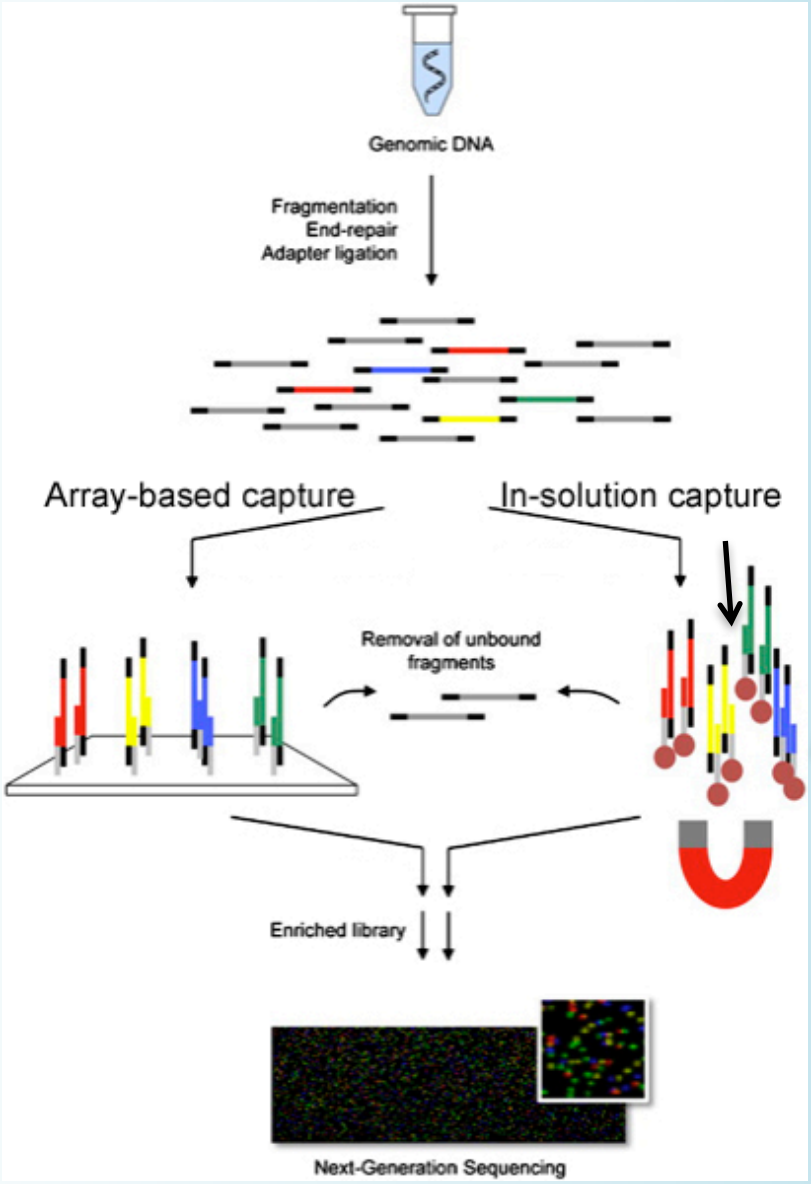
- Have a well-assembled reference genome
- Interested in non-coding regions
- Best way to look for signatures of selection since the vast majority of the genome will be covered
  - *For population genetic projects aiming at detecting natural selection we really need whole exomes to be targeted at least, and better yet , whole genomes*
- Have a reasonably small genome size or only need low coverage
- Have a larger budget or a smaller sample size
  - Ex. M. musculus low coverage (5x) WGS on HiSeq 4000 PE150 = 8 samples/lane, ~$350 each (including library prep)
  - However, buy-in cost is reasonably low
    - The expense is in the development of the genomic reference in the first place

What to do with genomic DNA libraries? Sequence the whole thing or capture a portion of the genome?

Capture enrichments

- Need a reference genome or well-annotated transcriptome
  - The latter can be obtained for a much cheaper cost than de novo WGS so long as there is properly-preserved tissue available for RNA extraction
  - After transcriptome assembly, annotation, and loci filtering, probes can be designed and synthesized
- Alternatively, some pre-designed kits are commercially available
  - But most are only for model organisms
- Mainly interested in particular genomic regions
- Have a large sample sizes and a moderate budget
  - Probe buy-in is in the $1000's (except for UCE's) but can be used for many samples
  - Works out to ~$50-125/sample

Hybridization-based target enrichment



60bp unlabeled DNA probes

55-180bp biotinlylated RNA/DNA probes

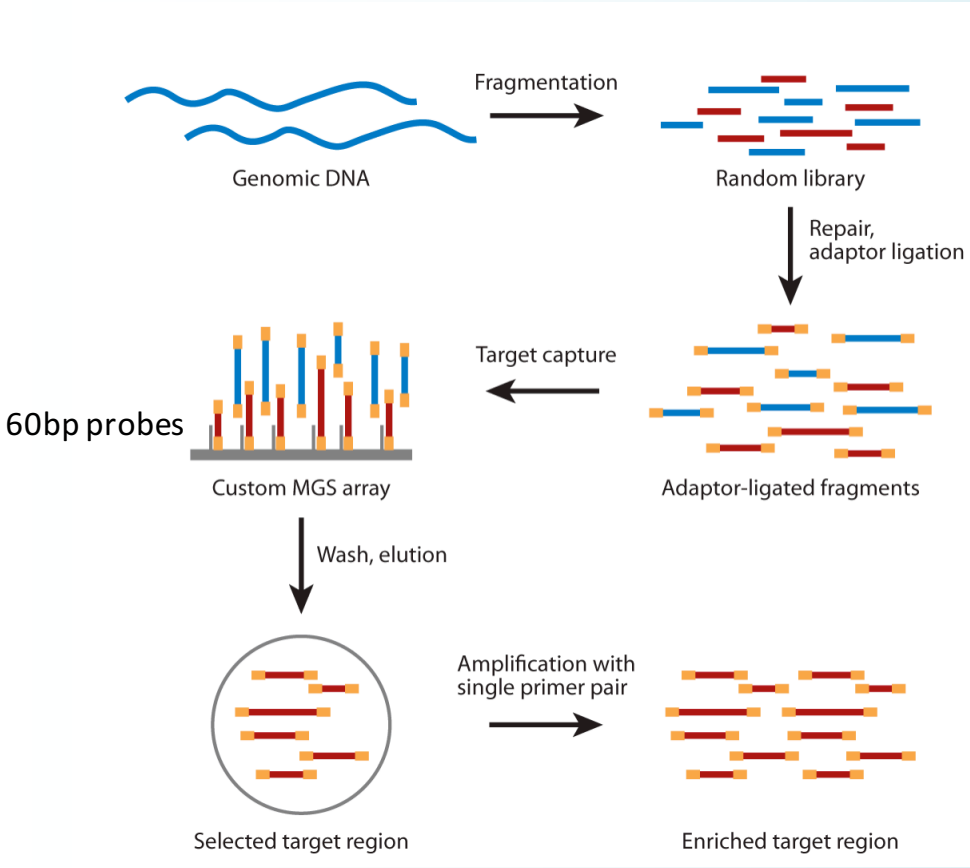
Baits synthesized by a commercial vendors (ie. Roche (Nimblegen), Agilent, Illumina, IDT, Mycroarray)

Target Enrichment – Microarray-based Capture

Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing

Emily Hodges[1,2], Michelle Rooks[1,2], Zhenyu Xuan[1], Arindam Bhattacharjee[3], D Benjamin Gordon[3], Leonardo Brizuela[3], W Richard McCombie[1] & Gregory J Hannon[1,2]

Nature Protocol 2009 4:960-974.



slide courtesy of Ke Bi

Agilent Custom SureSelect microarray 1M or 244K format

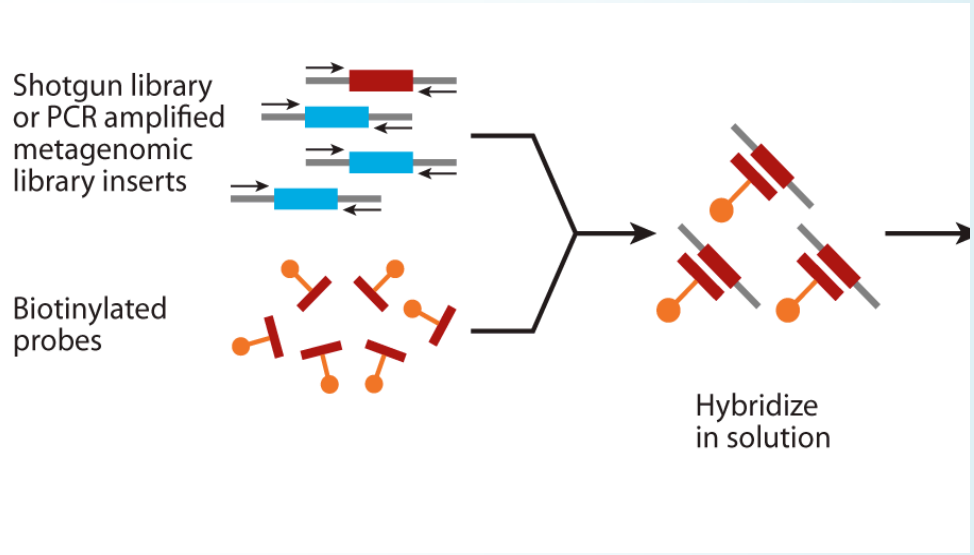
Pros

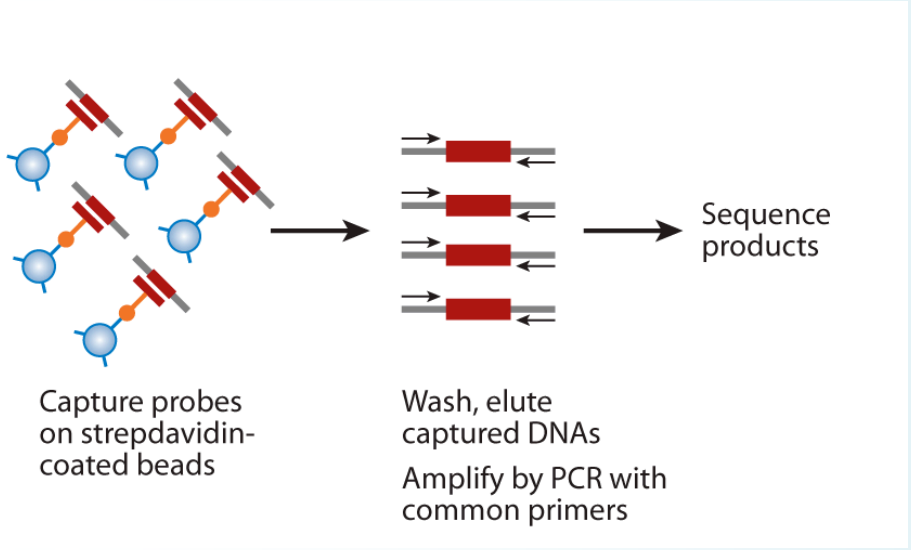
- Low cost: ~750USD for each 1M-probe array
- Suitable for small-scale phylogenomic and population genomic studies
- High probe tiling density/direct control over probe design

Cons

- Need a reference for probe design (also true for other types of hybridization-based methods.)
- Low capture efficiency
- Probe length short (60bp)
- Need special equipments (hybridization chamber, gasket slides, oven, etc. Available in EGL)
- Not cost-effective for surveying large number of samples
- Need large amount of input DNA (20µg/array) and Cot-1 DNA (50µg/µl)
- Complicated workflow

Target Enrichment - In-solution Capture





For non-model systems:

- NimbleGen SeqCap EZ Developer kits
- Mycroarray MyBaits kits

Pros

- Target size large (up to 200 Mb for NimbleGen) or medium (< 10 Mb) but with 16+ capture reactions per kit (MyBaits)
- Low amount of input DNA and Cot-1 DNA
- High level of multiplexing (Nimblegen > 50)
- Suitable for large-scale population genomic projects (NimbleGen) or phylogenomic projects (MyBait)
- High capture efficiency
- No special equipment needed)

Cons

- High initial investment (kits are more expensive than single array capture)
- Not cost-effective for multiple, small-scale population genomic projects when distinct target sets/design is required

slide courtesy of Ke Bi

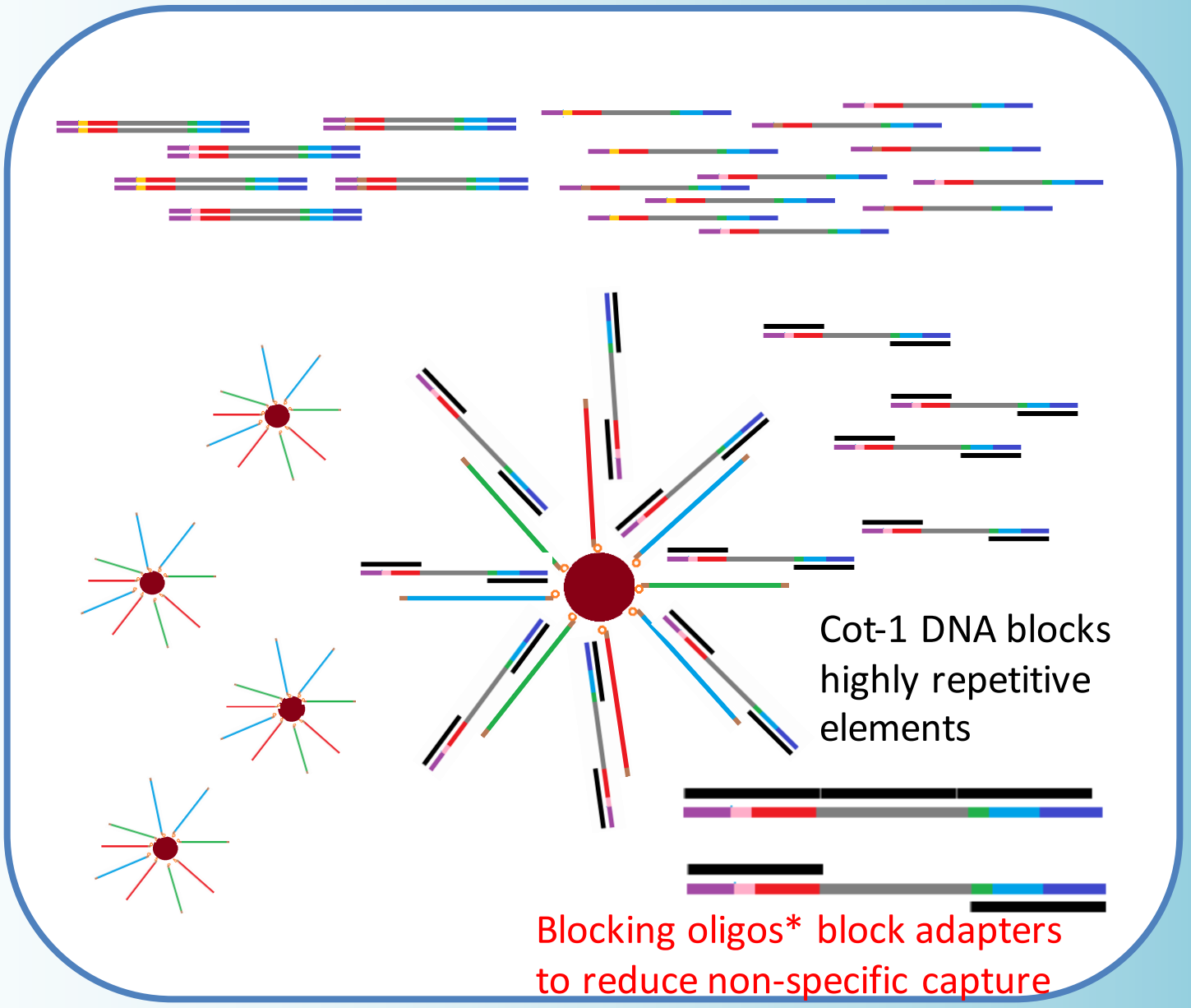
In-Solution Hybridization

Denature pooled libraries

Hybridize
With biotinylated probes and blockers
16-72 hrs at 47-65º C

Capture
Introduce streptavidin-coated beads. Pull probes & captured libraries from solution

***Blocking oligo selection is a key determinant of capture efficiency**



slide courtesy of Josh Peñalba

Nimblegen (Roche) SeqCap EZ Prime Developer Probes

Smallest custom kit sizes: (est. pricing includes buffer kit, sales tax & shipping)

- 4 reactions: ~$8000 (usually 5 reactions possible)
- 12 reactions: ~$13,000 (14-15 reactions possible)

The GSL and EGL both have standing quotes for Nimblegen kits.

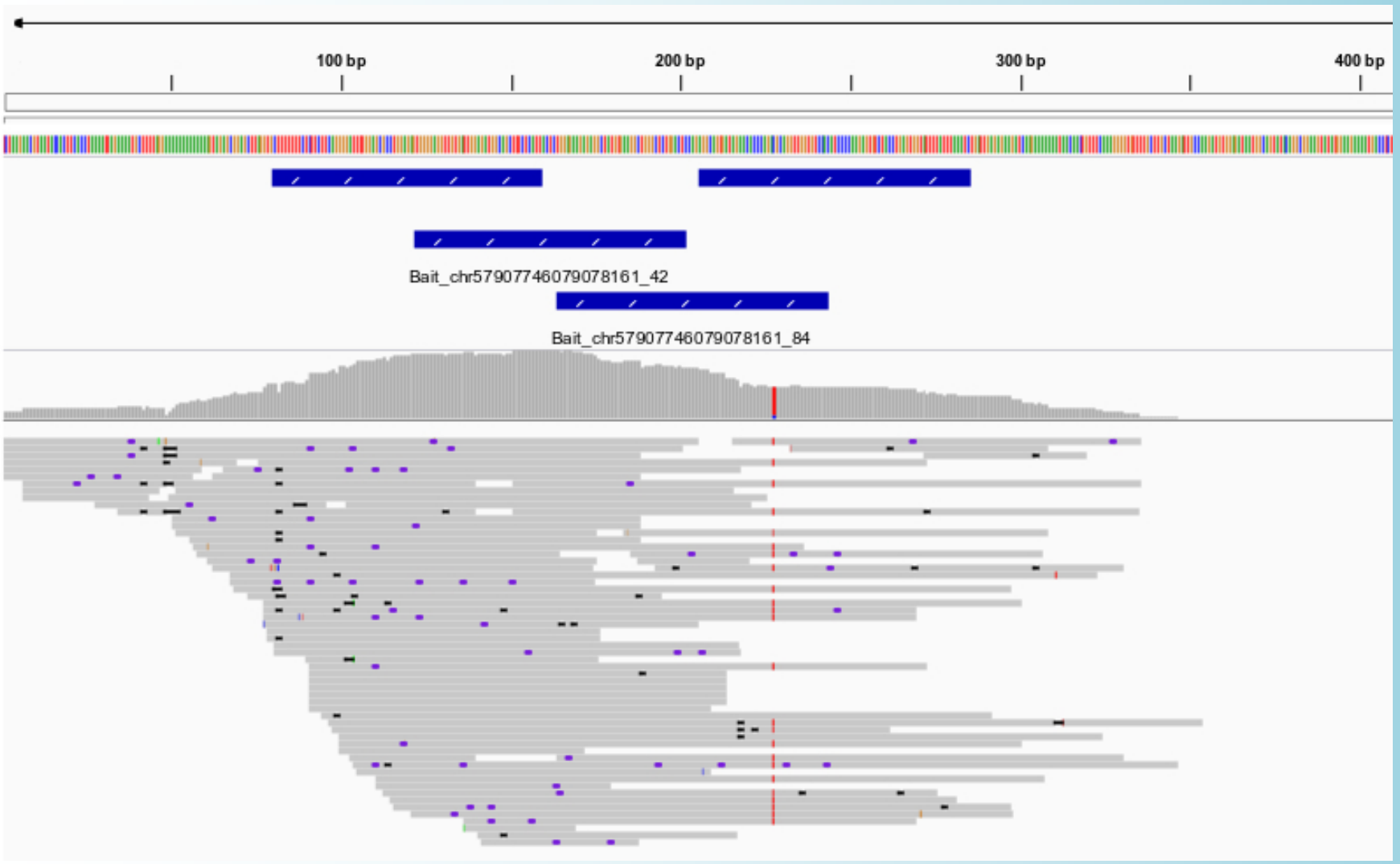
- GSL offers a capture service for $220 each reaction: you provide them with probes and a final library pool and they do the captures

Nimblegen SeqCap specs

- Up to 2.1 million unique probes in solution (1000's of copies of each probe), tiled as needed
- Total target size up to 200Mb
  - Typically whole vertebrate exomes are ~50 Mb
- Designed in consultation with Roche bioinformaticians from fasta files
- Multiplexing of large numbers of samples possible with that number of unique probes (20+)
- Pre-designed kits available for some model organisms

Nimblegen Probe Design

Example of probes with 2x tiling (blue bars)



Grey bars are sequence data that map to the reference

Nimblegen (Roche) SeqCap EZ Prime Developer Probes

Custom

- Custom sequence capture allows researchers to specify their own target regions of the genome.
- Researchers often use custom sequence capture as follow-up to Exome resequencing or CGH/SNP array studies.
- The opportunity for custom sequence capture spans from 100kb to 50Mb.
- Analyzing small target regions reduces resequencing costs even compared to exome sequence capture.

New (2017) improved manufacturing process: SeqCap EZ Prime		
Greater coverage of GC-rich target regions	Improved uniformity	Minimize the need for design refinements

Multiplexing for capture: What to consider

How many libraries can be captured together in the same reaction?

- What's your target size?
- What’s the expected genome size?
- How much tiling do the probes have?
- Desired depth of sequencing coverage?
- Sequencing platform and chemistry?
- Type of library prep?
- Library indexing and blocking oligo availability

- Check with bioinformatics specialists, probe manufacturers, and previous users of these methods for recommendations

Multiplexing for sequencing with capture data

- HiSeq 4000, PE150, 350M reads = 105 Gb data
- Target size = 5 Mb
- Desired average coverage = 50x
- Estimated percentage of on-target unique reads*** = 25%
  - 105,000,000,000 / 5,000,000 / 50 * .25 = 105 libraries multiplexed per sequencing lane

***This is the factor with the greatest uncertainty

Often ~50% with Nimblegen data but it is difficult to predict unless these probes have been used before

Expected on-target % (Specificity)

- Target size (easier to hit a larger target)
- Genome size (easier to accurately reduce a smaller genome)
- Library insert size (smaller libraries hybridize better)
- Blocking oligo choice (I recommend xGEN blocking oligos or individually-barcoded)
- Probe type
- Probe sequences (GC-rich regions more difficult to capture)
- Past experience of other researchers with similar probe sets and organisms/reported metrics in the literature

- Note: not all reads that map to targets are usable in alignments
  - Usually PCR duplicates are removed for analysis since they do not represent unique biological reads

Arbor Biosciences/Mycroarray MyBaits Custom Kits

Custom kit options start at $3600 (some discounts available)

Baitset Tier *(Max. # of bait sequences)*	Number of reactions			
	16	48	96	384
MYbaits-1 (20K)	$3,200	$5,760	$8,640	$23,040
MYbaits-2 (40K)	$4,000	$7,200	$10,800	$28,800
MYbaits-3 (60K)	$4,800	$8,640	$12,960	$34,560
MYbaits-4 (80K)	$5,600	$10,080	$15,120	$38,400
MYbaits-5 (100K)	$6,400	$11,520	$17,280	$46,080
MYbaits-6 (120K)	$7,040	$12,670	$19,000	$48,640
MYbaits-7 (140K)	$7,680	$13,820	$20,740	$53,760
MYbaits-8 (160K)	$8,320	$14,940	$22,460	$58,880
MYbaits-9 (180K)	$8,960	$16,130	$24,190	$64,000
MYbaits-10 (200K)	$9,600	$17,280	$25,920	$69,120

Kit item number: MYbaits-[# of modules]-[# of reactions]

- In-solution probes in modules of 20,000 designed from fasta files of targets
- Multiplexing is common, but in smaller numbers (< 20) due to fewer unique probes
- Very cost-effective for many captures
    - price per capture decreases as more reactions are ordered)
    - great choice for phylogenetics since more closely related species can be captured in separate reactions, reducing the risk of one library out-competing others

http://www.mycroarray.com/mybaits/mybaits-planning-your-project.html
http://www.mycroarray.com/mybaits/mybaits-custom-calculator.html

MyBaits methods papers

Special Issue: Sequence Capture

September 2016

Volume 16, Issue 5
Pages 1051–1278

Previous Issue | Next Issue

- **An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura) (pages 1069–1083)**
  Daniel M. Portik, Lydia L. Smith and Ke Bi
  Version of Record online: 24 JUN 2016 | DOI: 10.1111/1755-0998.12541
  Abstract | Article | PDF(949K) | References | Request Permissions

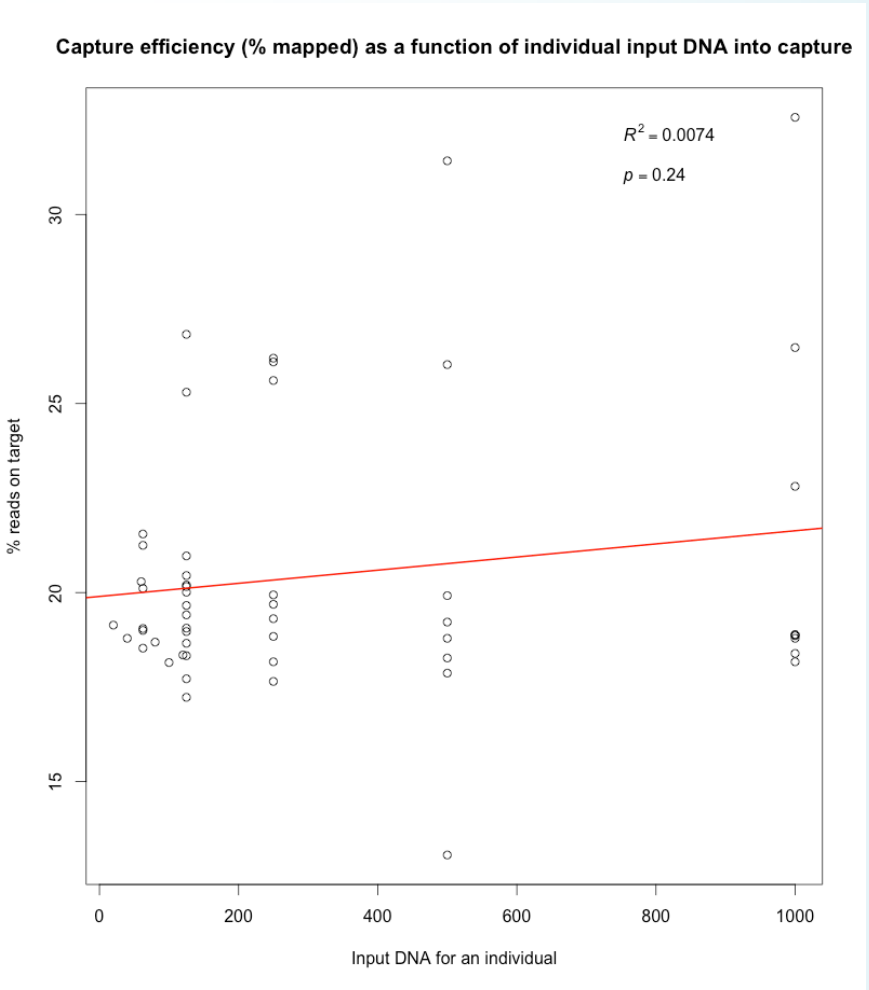
- **Exon capture optimization in amphibians with large genomes (pages 1084–1094)**
  Evan McCartney-Melstad, Genevieve G. Mount and H. Bradley Shaffer
  Version of Record online: 12 JUL 2016 | DOI: 10.1111/1755-0998.12538
  Abstract | Article | PDF(368K) | References | Request Permissions

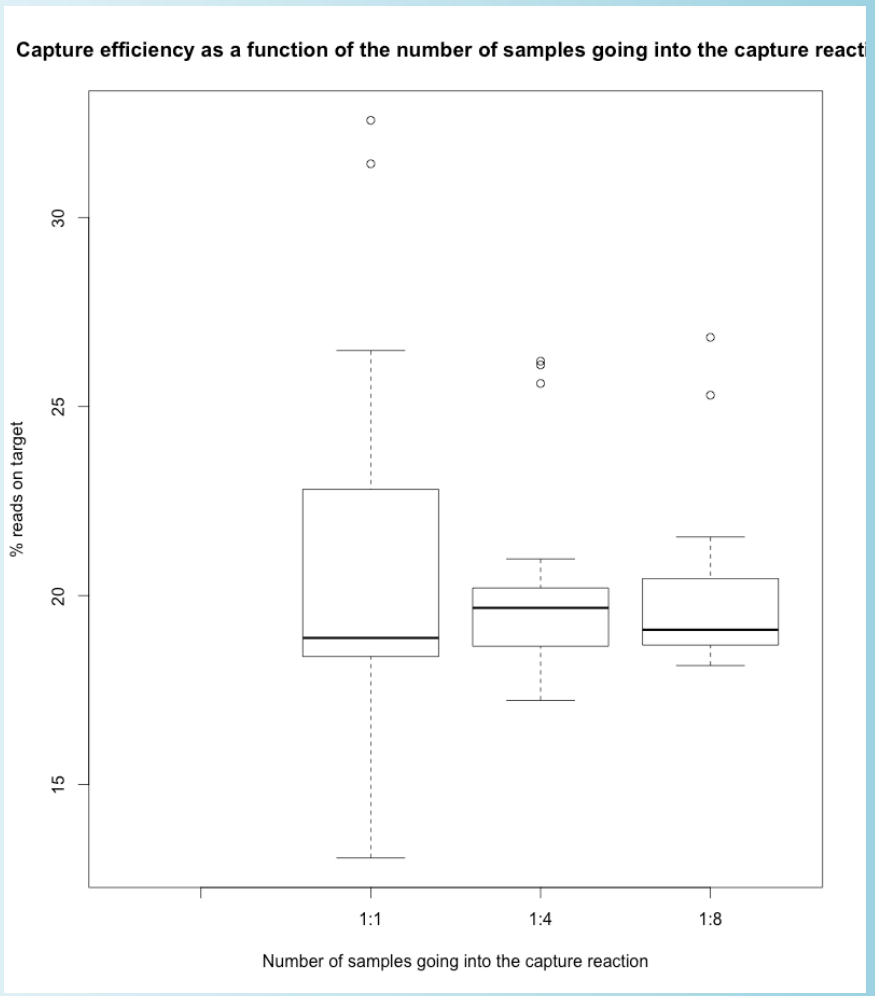
General cautions about literature

- But don't uncritically model a project after what you see in the literature.
- Publications tell you what is possible but not what is optimal
- A lot can change in the time from a project initiation → completion → publication → you reading about it
- Small changes to a project design can have large implications on outcomes
- Talk with experts, contact corresponding authors to see what the best practices are for today

General capture conclusions (MyBaits): capture efficiency

A slight decrease in capture efficiency (specificity) is seen with increase in number of individual libraries multiplexed in each reaction.

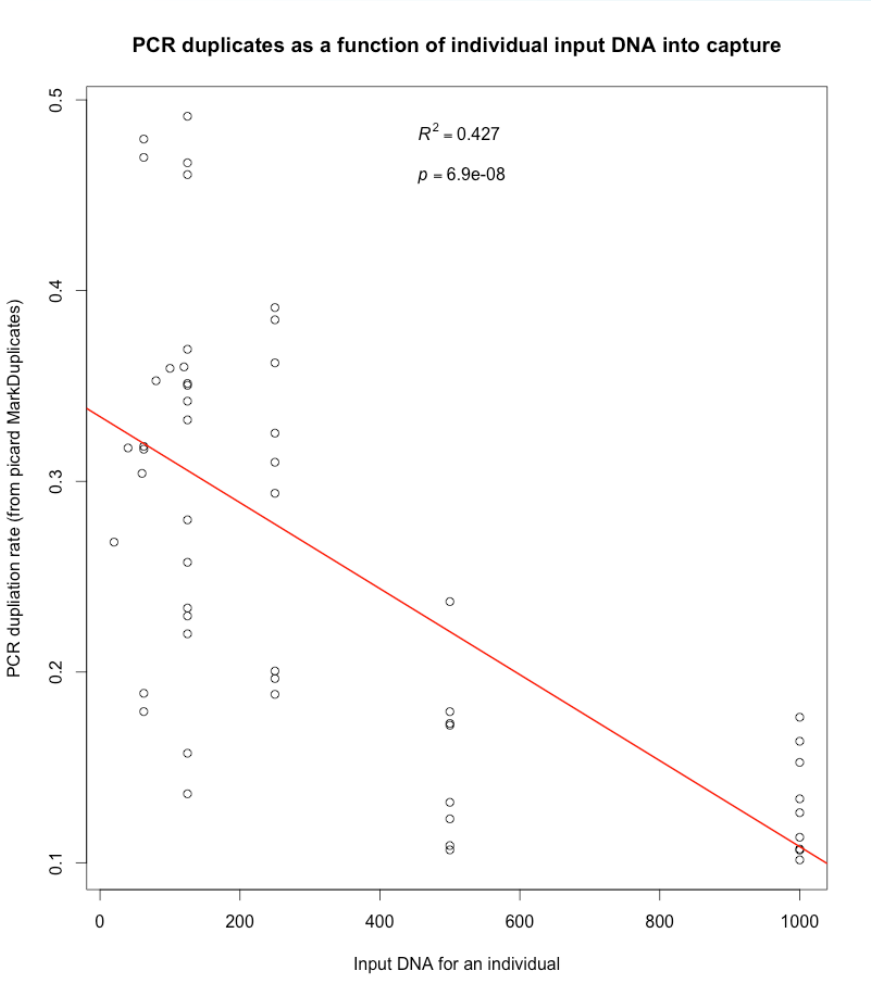




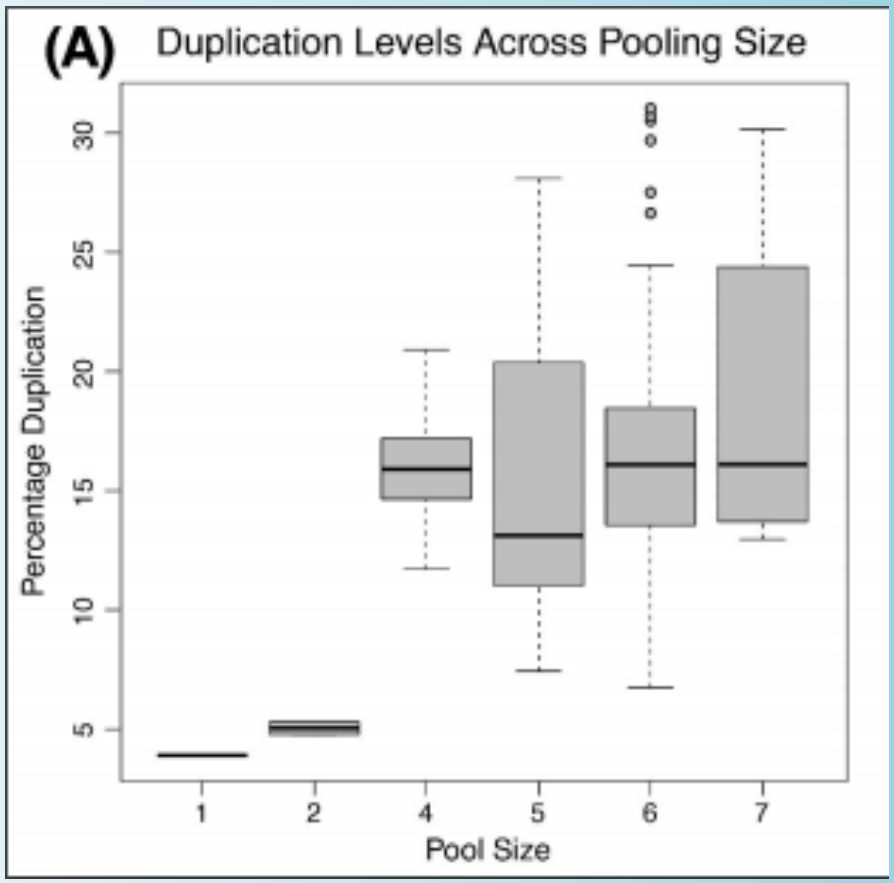
Evan McCartney-Melstad, et al. 2016

General capture conclusions (MyBaits): PCR duplicates

The proportion of duplicate reads increases with the increase in number of individual libraries multiplexed in each reaction.



Evan McCartney-Melstad, et al. 2016



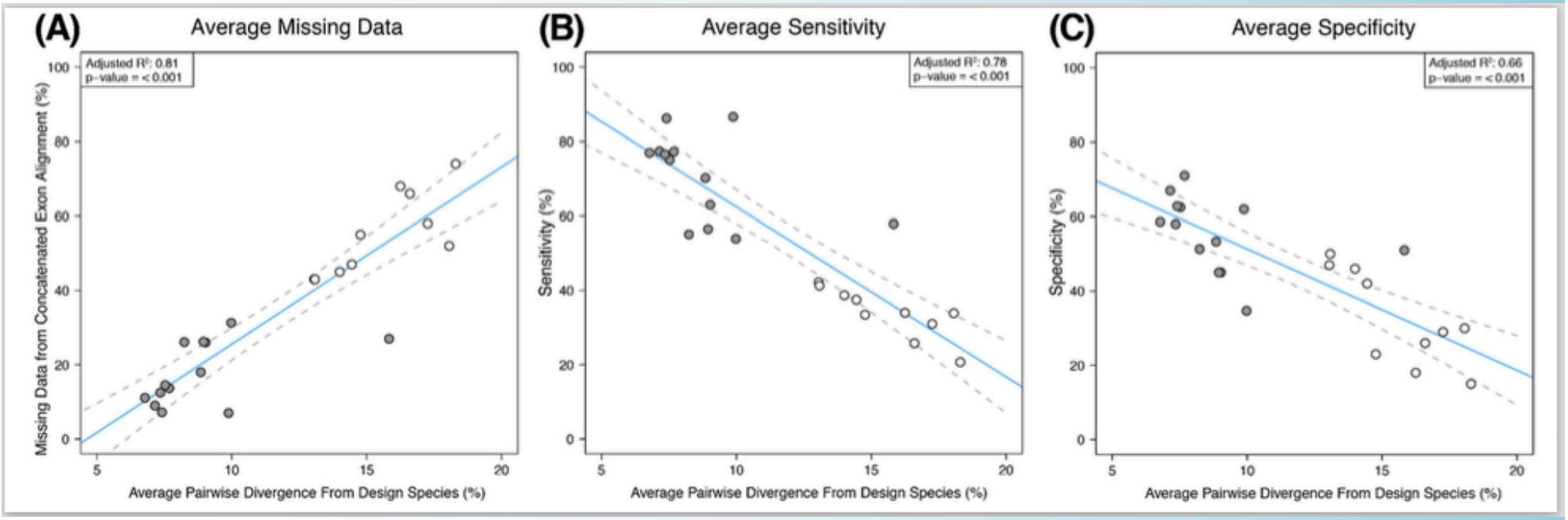
Portik, Smith & Bi

General capture conclusions (MyBaits): distance from probes

Utility of MyBaits probes for phylogenetic studies: ingroup samples up to 10% divergent (nuclear) capture quite well with 120mer RNA probes

- Ingroup samples (filled circles) up to 10% divergent (nuclear) capture quite well with 120mer RNA probes (share common ancestor ~56 mya), about 60% specificity
- Outgroup samples (open circles) 12-18% divergent (shared common ancestor ~77-103 MYA) are not as efficient, about 36% specificity

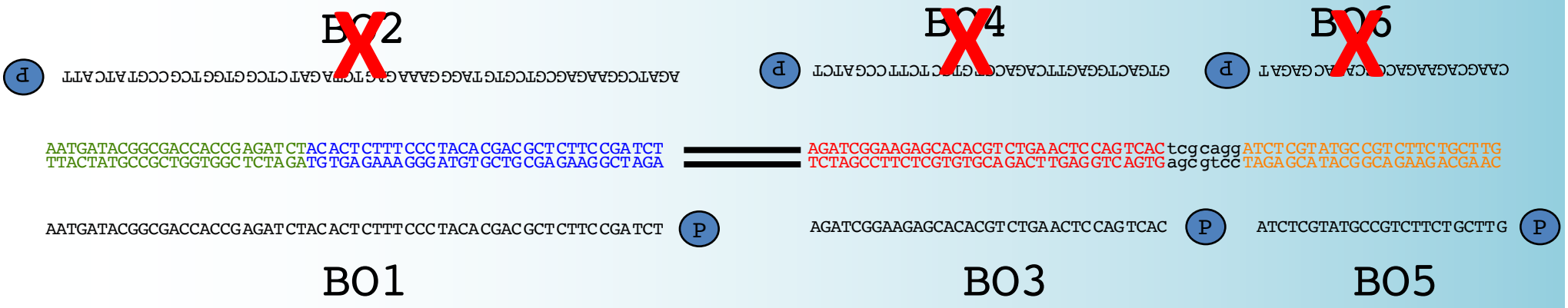
Specificity decreased 3.26% for each percent increase of pairwise divergence.



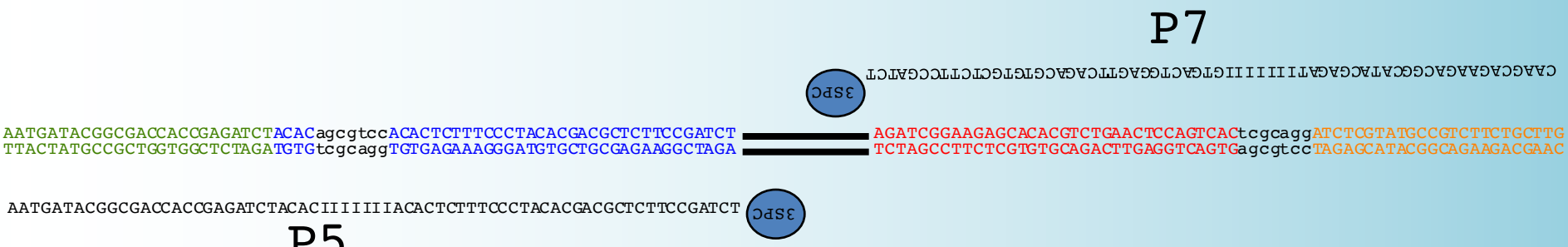
Portik, Smith & Bi

General capture conclusions (MyBaits): blocking oligo options

1) Short blocking oligos: Maricic, et al. 2010 and Meyer & Kircher 2010 (very cheap)

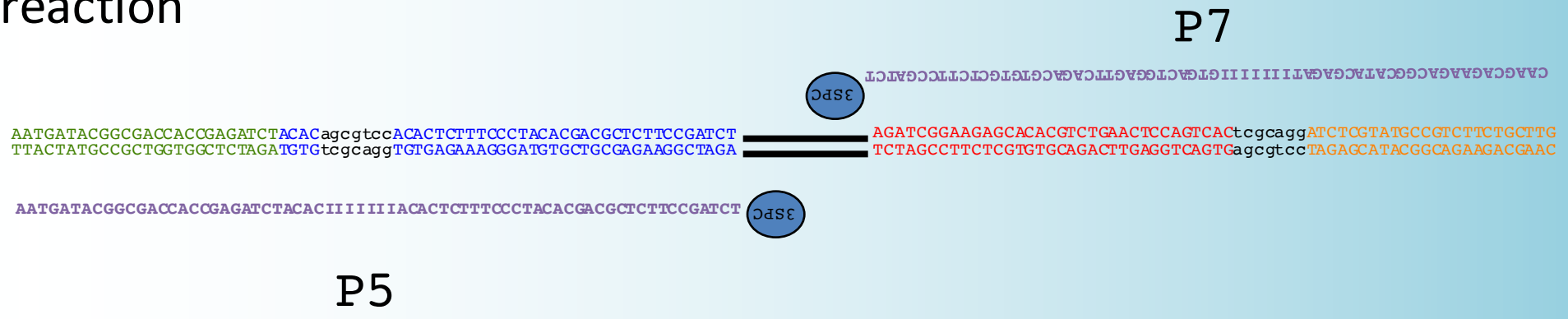


2) Generic blocking oligos: use inosine to block index(es): Come free with MyBaits kit (otherwise, cheap)

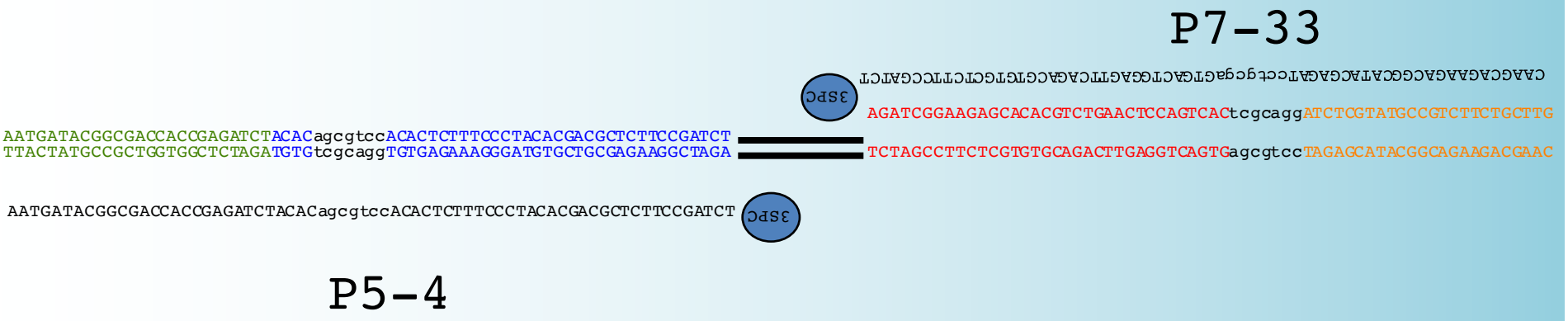


General capture conclusions (MyBaits): blocking oligo options

3) xGEN universal blocking oligos (IDT): use inosine to block index(es) plus special proprietary methylation mojo: $60+ per reaction



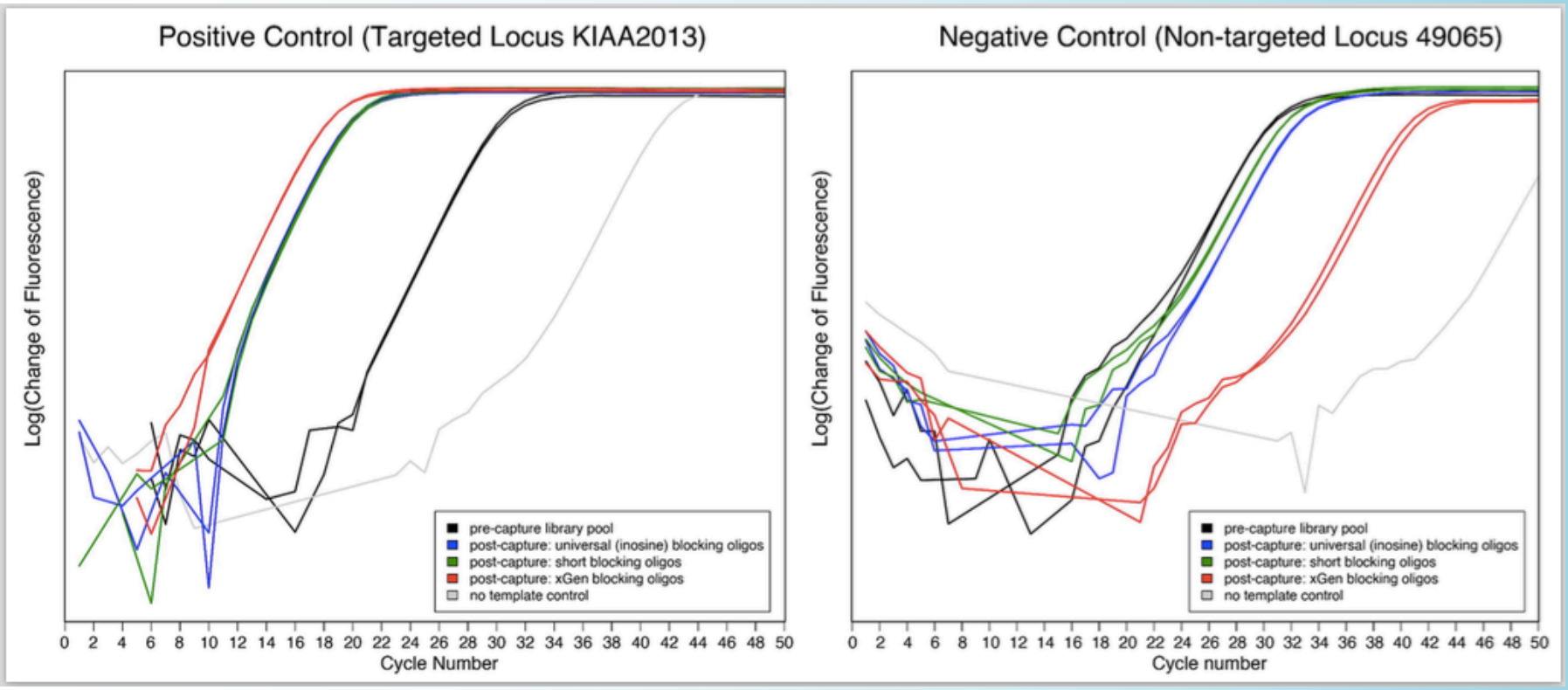
4) Individually-barcoded blocking oligos (IDT): exactly match the sequence of the index. Must have a full set available: $50/reaction



General capture conclusions (MyBaits): blocking oligos

A large increase in efficiency is seen in qPCR and sequencing results when using xGEN blocking oligos in comparison with short blockers or ones using inosine in the index (MyBaits)

Note: Individually-barcoded blocking oligos were not tested here



The further left the qPCR curve, the more copies of that locus are present in an equal amount of DNA

For a locus we targeted (left) the further left the curve, the better the enrichment

For a locus we did not target (right) the further right the curve, the better the depletion

Depletion is often the better signal of a more successful capture

Portik, Smith & Bi 2016

General capture conclusions (MyBaits): blocking oligos

IDT xGEN blocking oligos were tested against generic inosine blockers provided with MyBaits kit, and samples taken through sequencing

Note: Individually-barcoded blocking oligos were not tested here

Library	Specificity(%)	Sensitivity(%)	AvgCoverage(X)
MB_blocker_1	10.17	98.31	9.07
MB_blocker_2	10.66	98.76	10.54
MB_blocker_3	8.89	98.61	11.17
MB_blocker_4	10.83	98.95	12.18
MB_blocker_5	11.28	98.7	12.56
MB_blocker_6	9.72	98.92	13.44
MB_blocker_7	11.06	98.78	13.67
MB_blocker_8	10.94	98.93	14.27
MB_blocker_9	9.64	98.98	14.64
MB_blocker_10	10.05	98.97	14.76
XG_blocker_1	40.85	99.53	45.93
XG_blocker_2	42.68	99.62	47.61
XG_blocker_3	42.09	99.75	48.3
XG_blocker_4	40.91	99.69	56.19
XG_blocker_5	44.37	99.77	72.48
XG_blocker_6	45.77	99.91	73.37
XG_blocker_7	43.07	99.86	80.54
XG_blocker_8	45.25	99.79	83.86
XG_blocker_9	44.11	99.84	84.07

- xGEN blocking oligos cost ~$60 per capture
- But they resulted in 4x the specificity (mapping efficiency) and therefore 4x the coverage
- Far more money can be saved by reducing sequencing costs: worth it to pay $60 more per capture in order to use ¼ the number of sequencing lanes
- Note: conclusions made in Portik et al. about blocking oligo choices were cited before publication from biorxiv and saved at least one researcher precious research dollars.

Real world project cost examples: pilot MyBaits project (Portik, Smith & Bi, 2016)

DNA Library Preps and Captures	SUBTOTAL
264 DNA Library Preps	$4,266.89
Custom MYbaits-3 capture kit, 60k probes, 48 rxns	$4,731.89
Capture reaction costs	$4,504.91
Illumina HiSeq2500, 100 bp PE reads (3 lanes)	$7,614.00
Sequence-Capture Total	**$21,117.69**
Approximate cost per sample	**$79.99**

- Target size = ~1 Mb; 5-6 libraries/capture; 80-90 libraries/Illumina lane
- 68% of reads pass QC filters; of this 36% (outgroup) - 60% (ingroup) map to targets
- 142X average coverage!!! This was definitely oversequenced since we did not know what to expect
- Today, sequencing costs would be only 1/3 since the whole project could be run on HiSeq4000 PE 150 for $2450 and still have 100x coverage → $60/sample

Real world project cost examples: Nimblegen (large: rodent whole exome)

	Unit Price	Quantity	Total		Notes
Nimblegen 12 reaction kit	8161	1	**13229**		(Note: the 12 reaction kit will usually allow 14-15 reactions)
Illumina Sequencing HS 4000 PE150	2320	5	**11600**		Including qPCR for 5 capture pools
DNA Extraction	*2*	*255*	*510*		*approx cost for enzymes, alcohol, buffers, plastics, qubit, agarose gel*
Libraries	*20*	*255*	*5100*		*approx cost for reagents, plastics, agarose, qubit*
Capture assessments	*50*	*15*	*750*		*approx cost for PCR assessment, bioanalyzer, qubit, and pooling*
Capture reagents (non-probes)	*100*	*15*	*1500*		*approx cost for blocking oligos, streptavidin-coated beads, cot-1*
EGL charges total	7860	1	**7860**		Estimated EGL bill for the project
			32689		Total (USD)
			$128.19		Cost per sample (USD)

Target size = 52 Mb; 17 libraries/capture; 51 libraries/Illumina lane
(low coverage requirements, reference alignment)

Real world project cost examples: Nimblegen (small)

	Unit Price	Quantity	Total		Notes
Nimblegen 4 reaction kit	8161	1	**8161**		(Note: the 4 reaction kit will usually allow 5 reactions)
Illumina Sequencing HS 4000 PE150	2340	1	**2340**		Including qPCR for 5 capture pools
DNA Extraction	*2*	*200*	*400*		*approx cost for enzymes, alcohol, buffers, plastics, qubit, agarose gel*
Libraries	*20*	*200*	*4000*		*approx cost for reagents, plastics, agarose, qubit*
Capture assessments	*50*	*5*	*250*		*approx cost for PCR assessment, bioanalyzer, qubit, and pooling*
Capture reagents (non-probes)	*100*	*5*	*500*		*approx cost for blocking oligos, streptavidin-coated beads, cot-1*
EGL charges total	5185	1	**5185**		Actual EGL bill for the project
			15686		Total (USD)
			$78.43		Cost per sample (USD)

Target size = ~4.25 Mb; 40 libraries/capture; 200 libraries/Illumina lane (high coverage requirement)

Don’t be discouraged by these prices

- Both Nimblegen and Mycroarray sales reps know we are usually working on limited budgets compared with industry and biomedical research
- They would rather make a smaller sale than no sale at all and are willing to offer discounts or customize volumes for students and post-docs
- If capture hybridizations are a good fit for your project *scientifically* do whatever you can to make the budget component work rather than settling for a subpar method.
- Our sales reps are:

Emma Murphy (Nimblegen/Roche):

emma.murphy@roche.com

Alison Devault (MyBaits/Mycroarray/Arbor):
alison@mycroarray.com

Other probe vendors/methods

I highlighted MyBaits and Nimblegen two vendors that we use quite a bit since:

1) they are willing to discount pricing and work with relatively small budgets
2) they offer good design support for non-model organisms
3) we now have benchmark data and optimized protocols

But they are far from the only choices:

- Illumina
- Agilent SureSelect
- IDT
- DIY methods
  - SCPP: Peñalba, Joshua V., et al. "Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms." *Molecular Ecology Resources* 14.5 (2014): 1000-1010. DOI: 10.1111/1755-0998.12249
  - HyRAD: Suchan, Tomasz, et al. "Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens." *PloS one* 11.3 (2016): e0151651. https://doi.org/10.1371/journal.pone.0151651
  - HyRAD X: Schmid, Sarah, et al. "HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA." *Methods in Ecology and Evolution* (2017). DOI: 10.1111/2041-210X.12785

Predesigned Kits

If your design project allows, ordering pre-designed kits can be a less expensive way to undertake an in-solution targeted enrichment project.

- Nimblegen, Illumina, Agilent: human exon, UTR + exon,
- Nimblegen: exome capture (mouse, pig, canine, soy, wheat, barley, maize, switchgrass)
  - Note: Nimblegen is not off the shelf, same synthesis time as custom kits, but reduce (i.e. no) design time
- MyBaits: Whole Genome Enrichment
- MyBaits: Mitochondrial bait (human, many mammals, salmon)
- MyBaits: Ultaconserved Elements
  - Tetrapods
  - Ray-finned fish
  - Hymeoptera: general & ant-specific
  - Asteraceae (Compositae) flowering plants

Ultraconserved Elements (UCEs)

- Highly conserved regions of organismal genomes shared among evolutionary distant taxa
- UCE probes are anchors to capture adjacent genomic areas which are more variable
  - unlike other capture methods, we are mainly interested in the flanking regions here, not the targets
  - Usually require larger library inserts (~500 bp)
- Developed for resolving deep phylogenies but flanking sites may have enough variation for population genetic/phylogeographic analyses as well.
  - Leaché et al. 2015: Study using/comparing both RAD-markers and UCEs. doi:10.1093/gbe/evv026
- Great website (ultraconserved.org) with protocols, papers/talks, FAQs, and probe sequences
- Because probes are published, researcher can custom order a subset (as with Leaché et al.), include them with other capture probes, or order a pre-made kit from MyBaits

Target Enrichment - Ultraconserved Elements (UCEs) Capture

Syst. Biol. 61(5):717–726, 2012

DOI:10.1093/sysbio/sys004
Advance Access publication on January 9, 2012

Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales

BRANT C. FAIRCLOTH[1,*], JOHN E. MCCORMACK[2], NICHOLAS G. CRAWFORD[3], MICHAEL G. HARVEY[2,4], ROBB T. BRUMFIELD[2,4], AND TRAVIS C. GLENN[5]

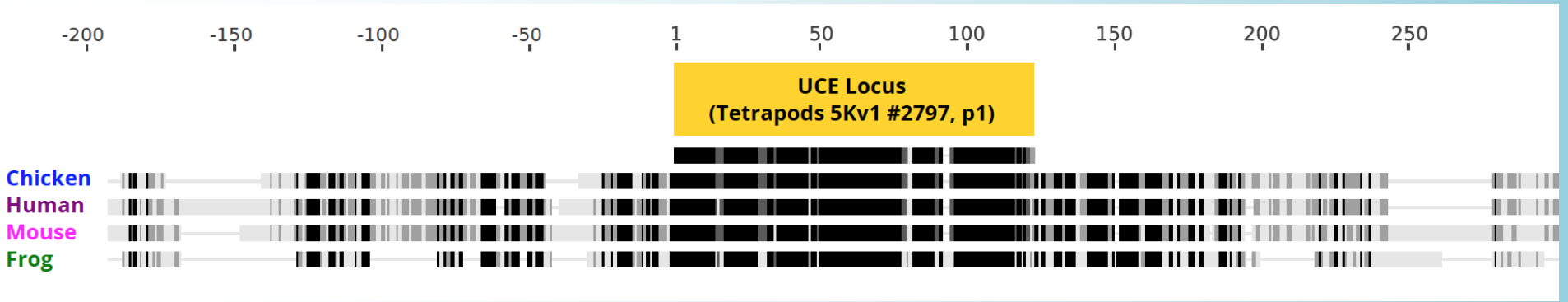
- Mycroarray MYbaits-UCEs kits
- Order probe sequences as part of a custom kit

Pros

- Cheap (e.g. 5K loci tetrapod kits cost about $700).
- No need for marker selection and probe design: 4K loci known to work for birds & reptiles, 2-3k loci in mammals, and up to 1k loci in amphibians
- Shown to be robust for resolving different phylogenetic scales

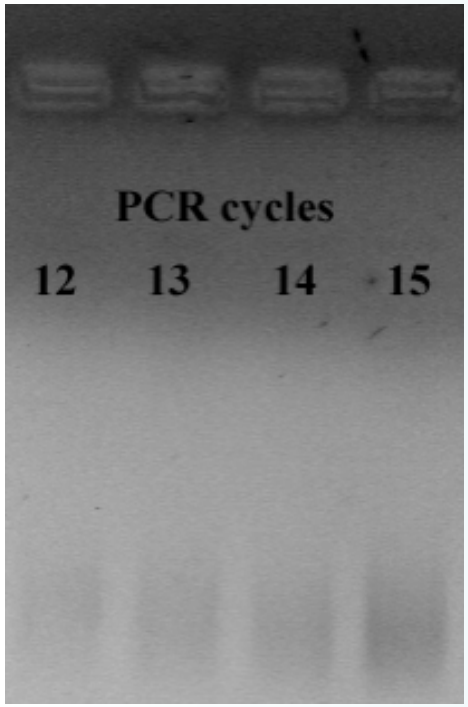
Cons

- Might not work well for heavily degraded samples (historic DNA) since it requires genomic libraries with relatively large inserts (>500bp)
- Not necessarily a good choice for intra-specific studies
- Limited to already developed probe sets



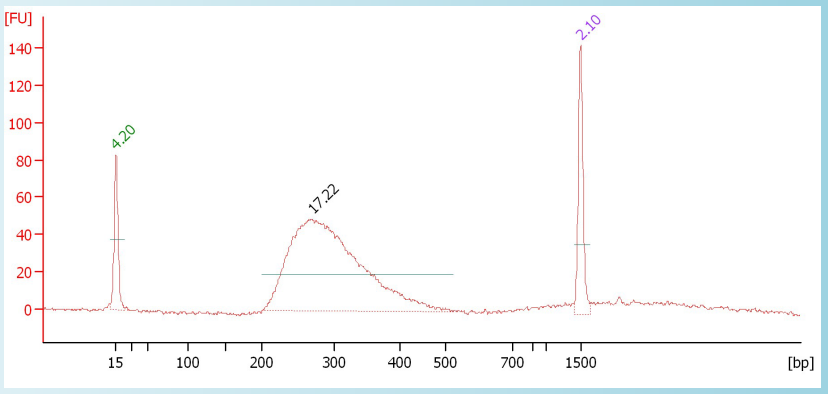
Post Capture Enrichment PCR – Avoid Over-amplification!

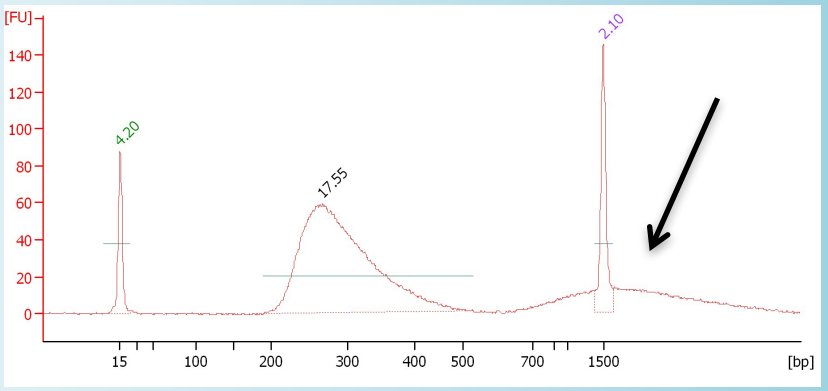
- In post capture enrichment PCR, there is a high probability of barcode swapping especially after PCR reaches saturation – short adapters can act as primers that may anneal to adapters containing different barcodes.
- Solution: amplify as few cycles as possible and never let your PCR reach plateau.
- To figure out how many cycles are needed – qPCR or quick PCR tests.



Concentration readings:
12 cycles: 12ng/ul
13 cycles: 19ng/ul
14 cycles: 28ng/ul
15 cycles: 35ng/ul

choose 12 or 13 cycles





Secondary hill = sign of overamplification

slide courtesy of Ke Bi

Post-capture enrichment PCR

Post-capture PCR involves a template where more than one library is present in the pool:

- If the reaction reaches plateau and runs out of available primer, adapters can prime each other to create chimeric molecules
- In addition, excessive numbers of PCR cycles will create an abundance of PCR duplicates in the final data
- Best to use fewer cycles but set up multiple reactions

Balance is needed between:

- Minimizing the number of PCR cycles
- Obtaining enough material for assessment and sequencing

For greater yields, PCR reaction with the DNA still attached to the beads: use Kapa reagents, Phusion not recommended

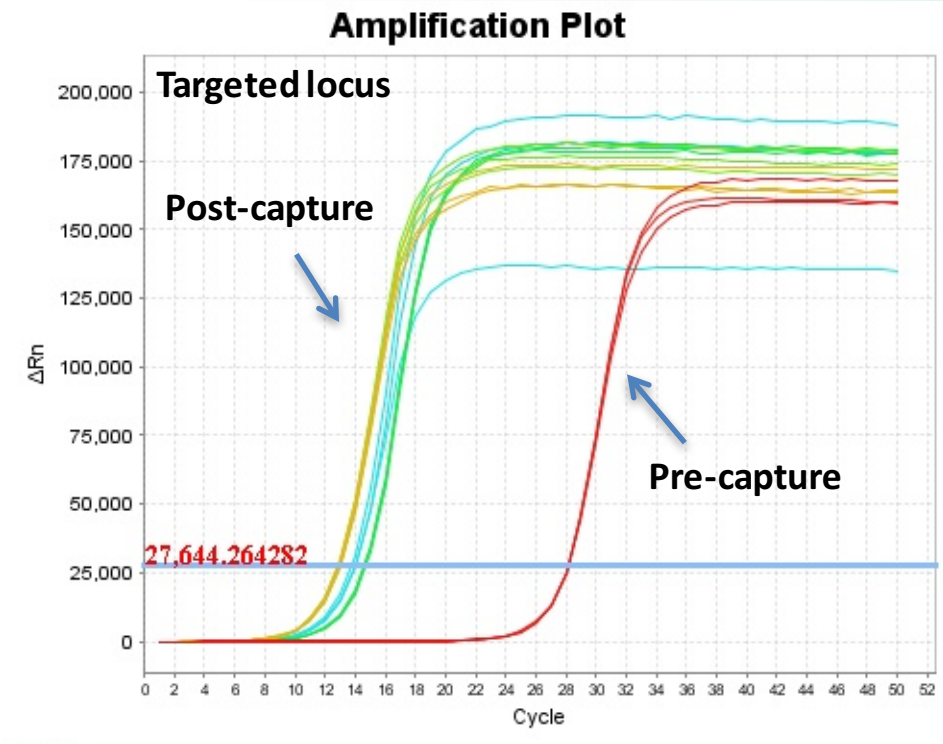
Berkeley sequencing (QB3) asks for at least 10µL of 10 nM concentration (= 0.1 picomoles) http://www.promega.com/a/apps/biomath/

Library averaging 500bp, 0.1 pmol = 33 ng of DNA (total)

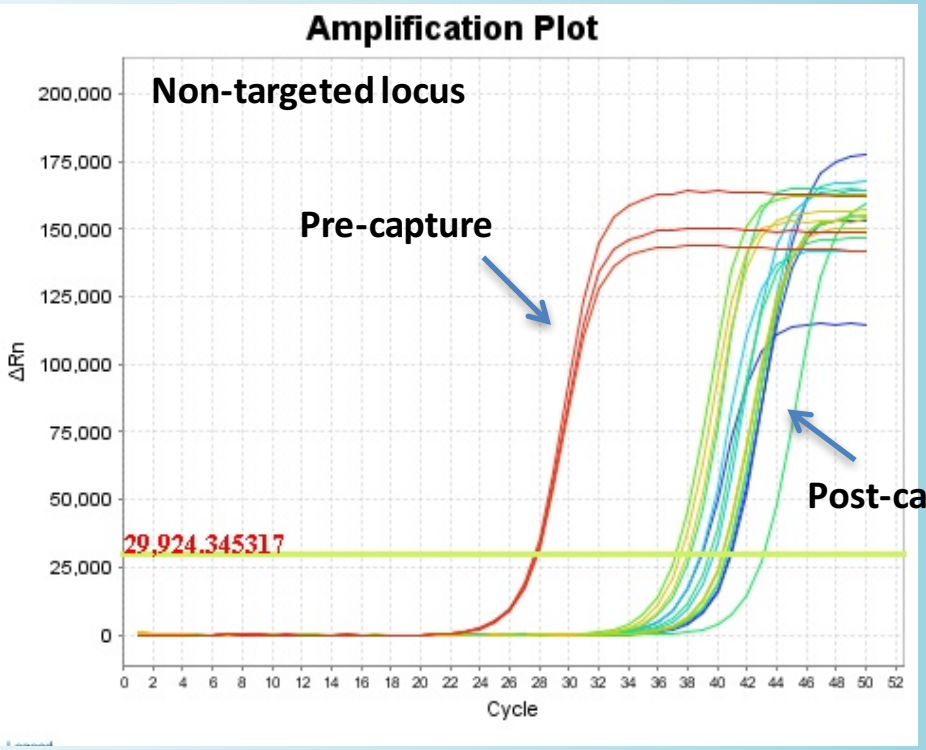
So, if there will be multiple reactions, even qubit results in the low single digits will be more than enough for sequencing.

Estimating Enrichment Efficiency using qPCR

qPCR assays are used to estimate relative fold enrichment by measuring the relative abundance of target loci (positive assays) and non-target loci (negative assays) in pre-capture sample library and post-capture captured multiplex DNA. These assays are an inexpensive way to determine whether the capture was successful prior to sequencing. The can't necessarily tell us how well a capture worked (Ct value changes and capture efficiency don't perfectly correlate). But a good qPCR assay indicates that a capture is worth sequencing.



Positive assay: enrichment of targets



Negative assay: depletion of targets

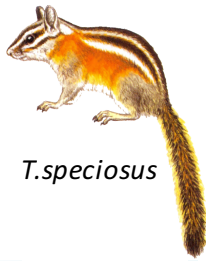
Performance of Transcriptome-based Exon Capture in a Case Study (Solution-based)

Total target size: 9.32 Mb

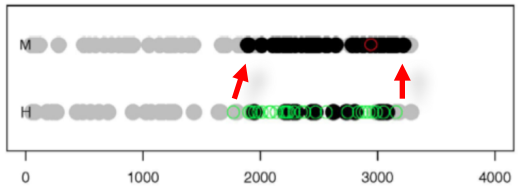
- ~2000 "candidate loci" in relevant pathways;
- 9774 assembled contigs with baits extended to their flanking regions;
- Control loci for contamination and qPCR.

Samples to survey: N = 303 + outgroups

Stable at Yosemite



T.speciosus

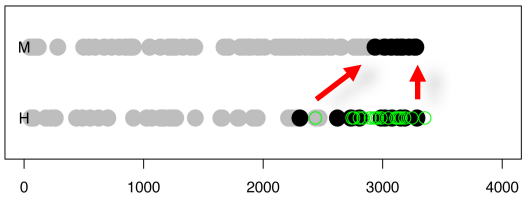


Modern: N=48

Historic: N=56

Retracting at Yosemite

T. alpinus

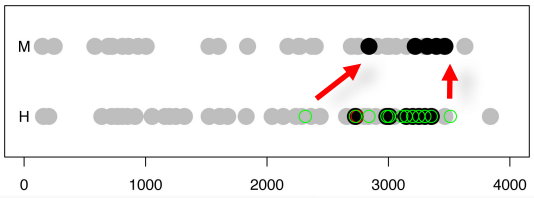


Modern: N=48

Historic: N=55

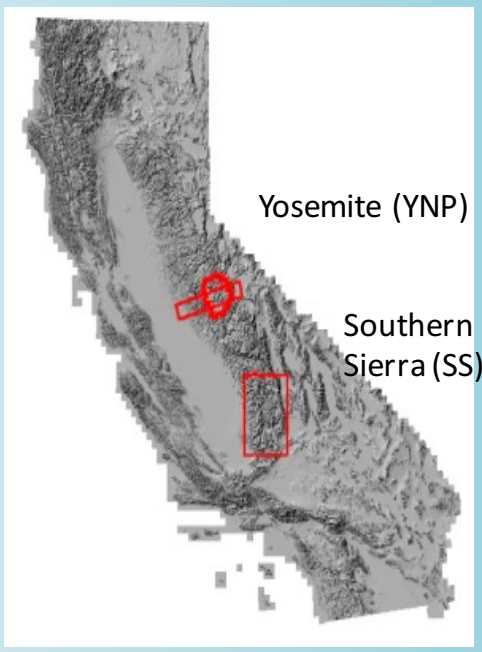
Retracting at southern Sierra

T. alpinus



Modern: N=41

Historic: N=55

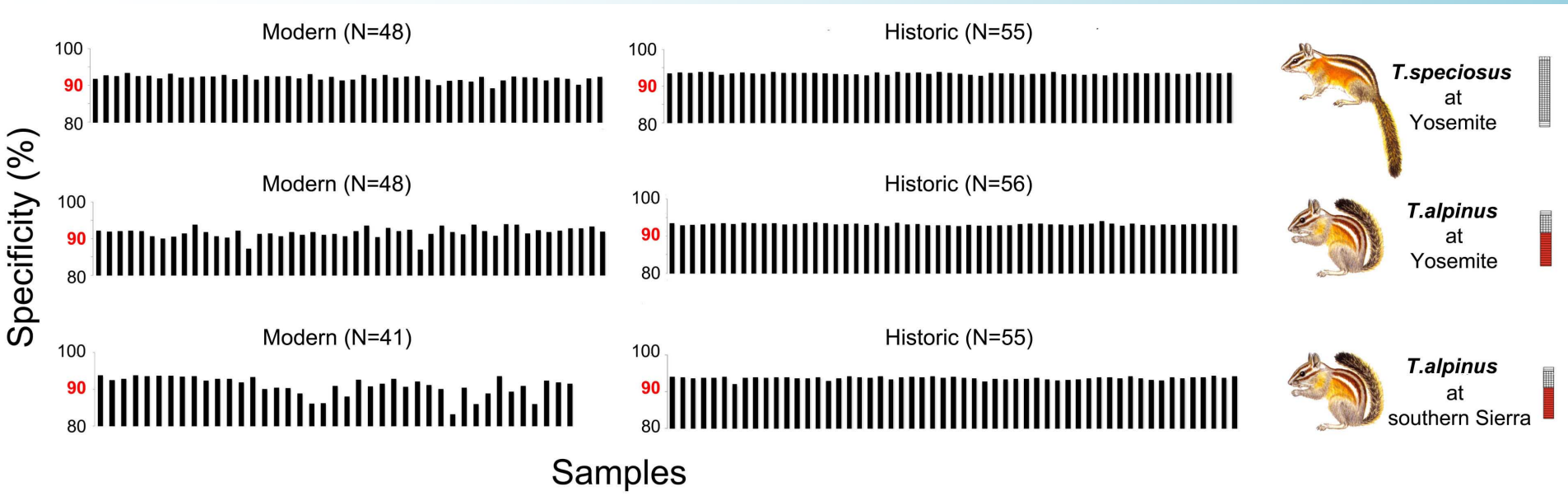


NimbleGen in-solution capture & sequencing:

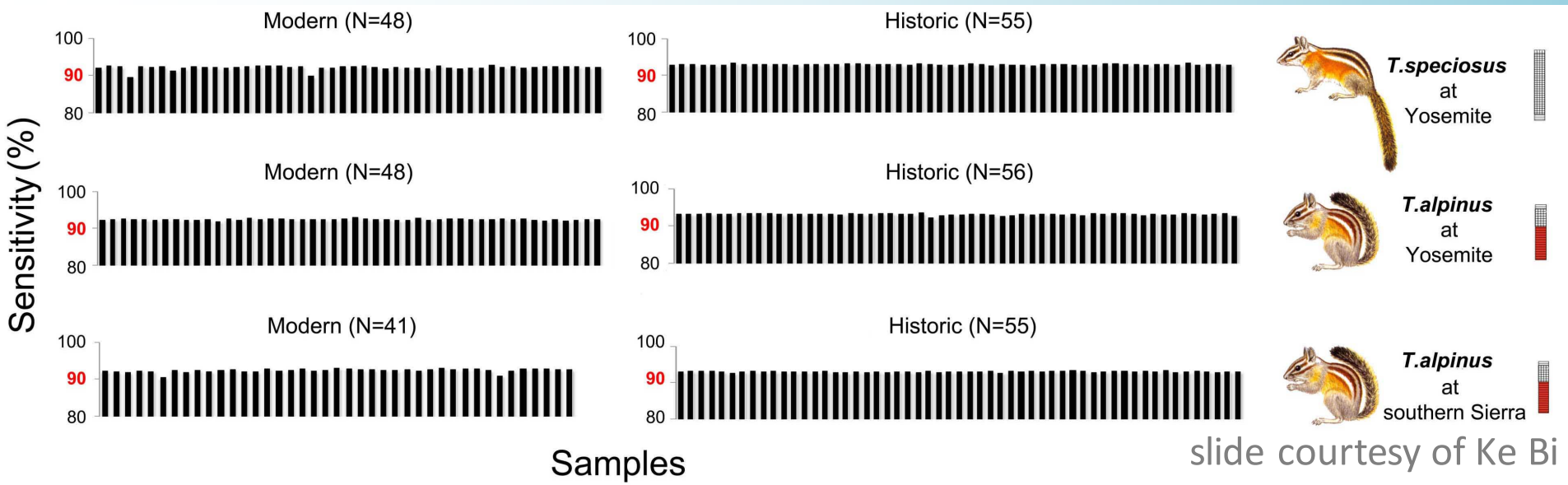
- Six capture reactions: 1 population/reaction;
- Illumina HiSeq2000, 100PE, 6 lanes: 1 population/lane.

slide courtesy of Ke Bi

Specificity - % cleaned reads mapped to the intended exons



Sensitivity - % target exons represented by sequence reads



slide courtesy of Ke Bi

Nimblegen captures: modern vs. historical samples

Recent Nimblegen capture metrics using:

Modern samples (from tissue collection)

- specificity = 62%
- data passing bioinformatics filters = 80%
  - 11% of reads are duplicates
- Average coverage = 12x

Historical samples (mammal skins)

- specificity = 75%
- data passing bioinformatics filters = 37%
  - 38% of reads are duplicates
  - Much higher adapter contamination (both read through of short inserts and residual adapters)
- Average coverage = 4x

Because historical libraries were grouped separately from modern ones for capture, the historical libraries could be sequenced on additional lanes in order to increase their coverage.

Performance of Nimblegen Capture in 4 Species of Hummingbirds

Testing for positive selection at candidate genes for high altitude adaptation while accounting for population history.

- **Approximate genome size:** 1Gb
- **Target size:** 6.6Mb (for candidate genes and random markers)
- **Approximate divergence time:** total divergence between all 4 species is ~20Ma
- **Total libraries captured & sequenced:** 233 (all from tissue collections)
- **Libraries pooled per capture:** up to 48 samples/capture reaction for total of 5 capture reactions
- **Total Illumina lanes and run type:** 2 lanes of HiSeq4000, 150bp paired-end.
  - All samples were run on both lanes and later combined for analysis
- **Average data retained after filtering:** 65%
- **Average specificity:** 75%
- **Average sensitivity:** 99%
- **Average coverage:** 64X (98% of sites have 5X coverage, 91% of sites have 20X)

In-solution Hybridization Methods

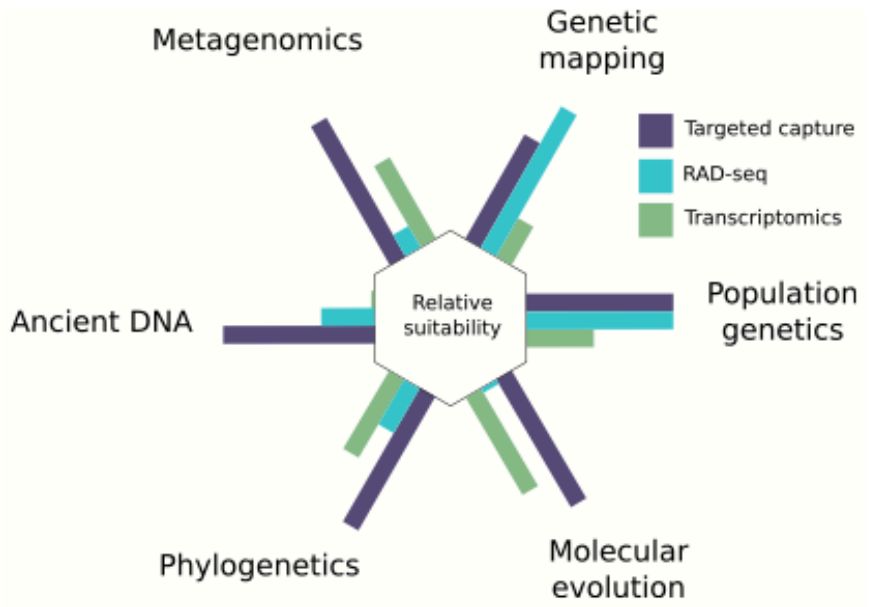


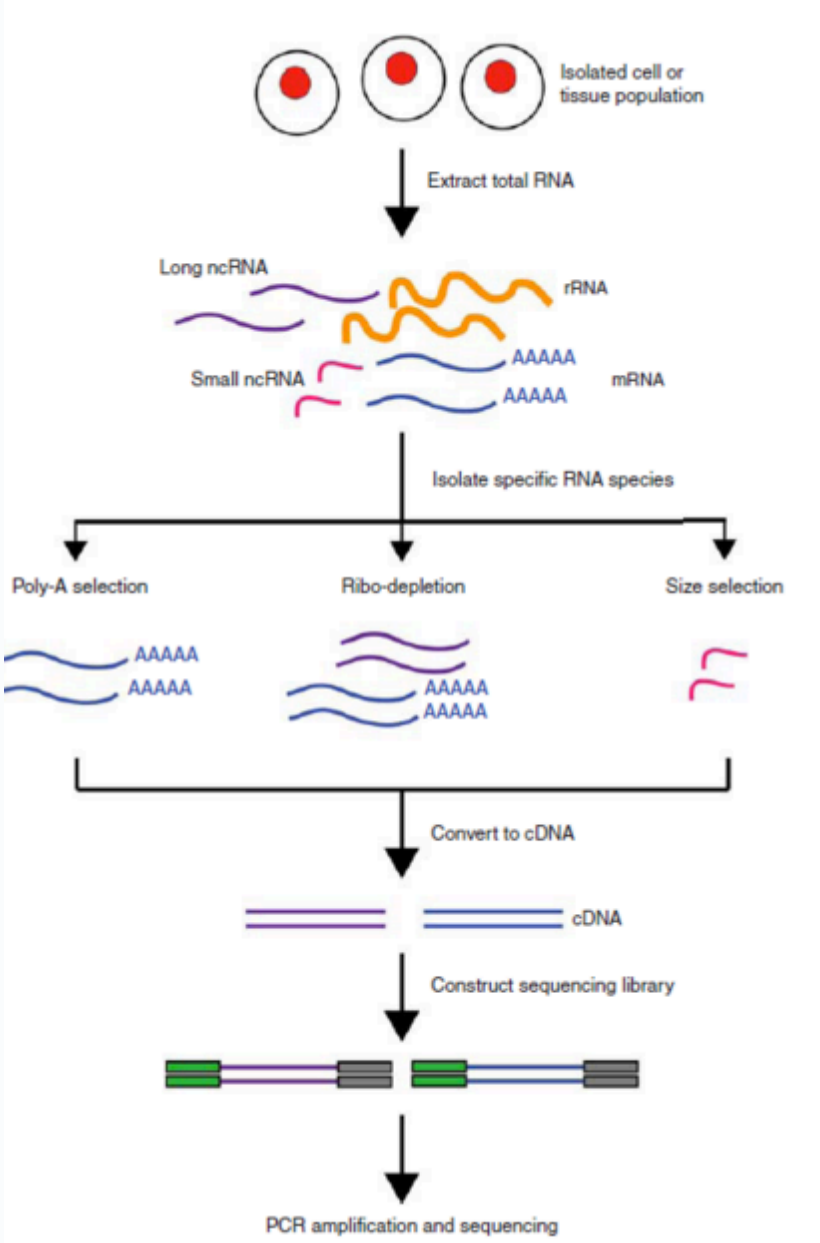
Fig. 4 The relative suitability of targeted capture (purple), RAD-seq (blue) and transcriptomic approaches (green) for genetic mapping of phenotypic traits, population genetics (includes inferring population history and detecting population-level signatures of selection), molecular evolution (e.g. rates of protein evolution), phylogenetics, ancient DNA and metagenomics. Specific height of each bar is arbitrary, but increasing height corresponds to increasing suitability.

Jones & Good, 2016
Targeted capture in evolutionary and ecological genomics

Molecular Ecology
doi: 10.1111/mec.13304

See also:
http://cgrlucb.wikispaces.com/file/view/Applications+of+Sequence+Captures+in+Non-model+Organisms.pdf

Illumina library preparation: RNA libraries



Kurkurba & Montgomery, 2015
doi: 10.1101/pdb.top084970

Illumina library preparation: mRNA isolation
1) Poly-A selection

- Most RNA-based projects (not all!) only interested in sequencing **mRNA** transcripts for differential gene expression or exon sequences
- 3' poly-A tail of eukaryotic mRNA can be captured with magnetic beads attached to oligo dT probes
- Relatively inexpensive approach; standard part of most RNA library kits
- Excludes all non-mRNA RNA. However, will also exclude transcripts with degraded poly-A tails

RNA Libraries: fragmentation

Use a few repeatable (or expendable) test samples to determine amount of time for the fragmentation step.
--mRNA must be turned into a cDNA library for sizing to be assessed
The Kapa protocol gives a range based on desired fragment size which tends to be fairly accurate, at least for animals.

Table 1. Recommended fragmentation conditions

Input RNA type	Desired library insert size (bp)	Fragmentation
Intact	100 – 200	8 min at 94°C
	200 – 300	6 min at 94°C
Partially degraded	100 – 300	1 – 6 min at 85°C
Degraded	100 – 200	1 min at 65°C

- But it can be hard to determine how much RNA is degraded. As a rule of thumb, for SR or PE100, we say:
  For RNA with a RIN score > 8, fragment for 6 minutes (at 94°C)
  For RNA with a RIN score < 8, fragment for 4 or 5 minutes (at 94°C)
- This is a starting point for testing only and is not a guarantee
- For PE150, larger inserts may be desired to maximize sequencing efficiency.

Illumina library preparation: mRNA isolation 2) rRNA depletion

- Samples that are degraded may no longer have a viable Poly-A tail for capture
- Samples in which sequence from all non-ribosomal RNA is desired (small RNA, viral RNA, noncoding regulatory RNA)

- RNaseH digestion of rRNA hybridized to probes (human/mouse/rat)
  - NEBNext rRNA Depletion Kit
  - Kapa Biosystems RiboErase Stranded Library Prep kit

- rRNA depletion with magnetic beads
  - RiboMinus (Thermo Fisher) (human/mouse, yeast, bacteria, plant, eukaryote)
  - RiboZero (Illumina): kit alone or library prep (human/mouse/rat*, bacteria, plant, yeast)
    - http://www.illumina.com/products/rrna-globin-mrna-removal-kit-selection-guide.html
    - *http://www.illumina.com/products/rrna-removal-kit-species-compatibility.html

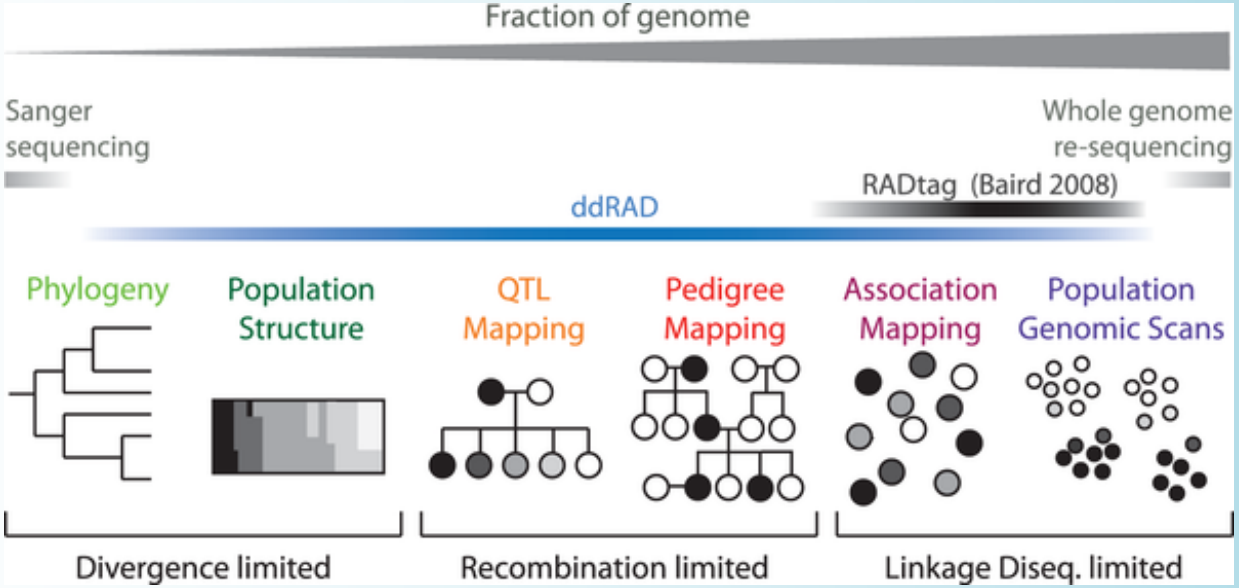
- Can be up to $100 per sample! (Just for rRNA removal: does not include extraction or library prep costs) But it is well worth it when there is no other way to obtain the RNA of concern.

RNA library best practices

- Proper RNA handling is vital prior to second strand synthesis. (After that, it is dsDNA and much heartier.)
- Fragmentation step may not be exactly the same as in the kit protocol (8 minute standard time is often far too long)
- Once RNA has been converted into cDNA, the same library process can take place as for genomic DNA, post-fragmention.
  - The only major difference is that starting material will be very low. Even greater care is required in handling and more PCR cycles may be required to have sufficient material for sequencing (often 10-15)
- Possible to split reagents
  - We always do so in the EGLwith RNA-seq for marker development or comparative sequences in the (saves 40% of library cost). Could likely reduce further if only common transcripts are desired.

Illumina library preparation: Restriction site Associated DNA

- RAD-Seq can uncover hundreds or thousands of polymorphic genetic markers across the genome in a single, relatively easy & fast, cost-effective experiment without any available reference genome
- Widely used for inferring population structure, phylogeography, lower level phylogenies, introgression, trait mapping… Many case studies available



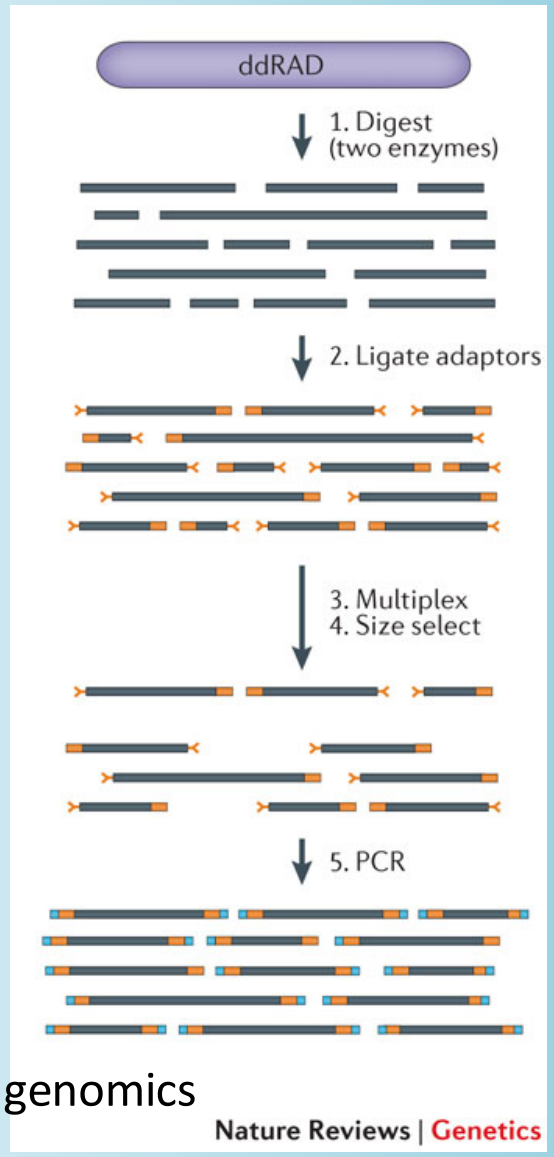
DOI: 10.1371/journal.pone.0037135

Illumina library preparation: RAD-Seq sounds amazing! what's the catch?

- Impossible to remove PCR duplicates with the most popular approach (ddRAD)
  - must do PE sequencing of single-digest RAD
  - Or use unique molecular indexes (UMI) as a part of the ddRAD adapter
- Can take a fair amount of time to select correct restriction enzymes (RE) and to optimize.
  - Simulation and outcomes do not always match
- Not appropriate for larger phylogenetic distances due to possibility of mutations at the RE cut sites
- Big problems with allelic drop-out and high variance in coverage depth across alleles and individuals.

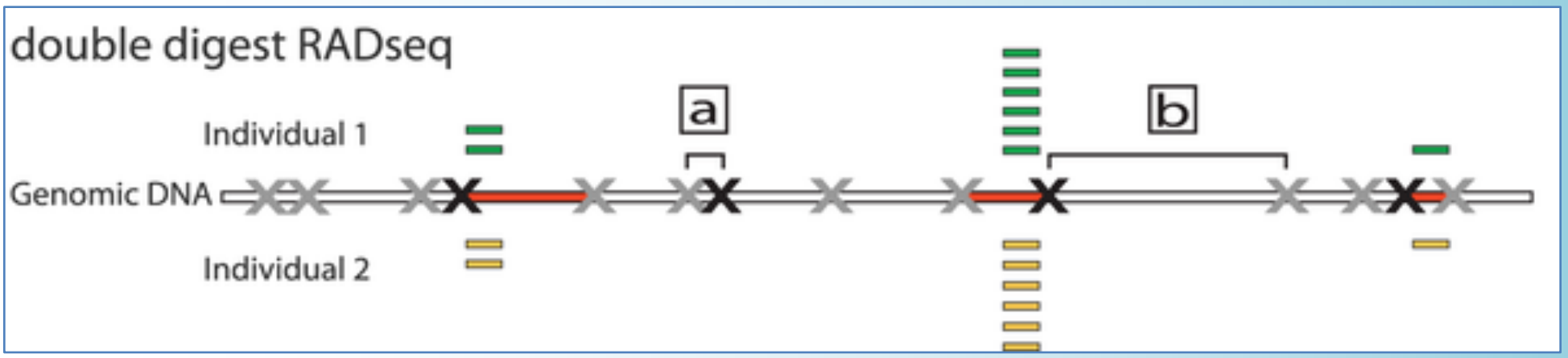
Illumina library preparation: ddRAD

- DNA is digested by two enzymes
- Adapters have overhang matching RE cut sites and internal barcodes
- Libraries are pooled and size selected with an automated instrument (Pippin Prep or similar)
- PCR amplification incorporates index and outer adapter (as in two-stage gDNA library) and enriches for fragments with the correct inner adapter combination



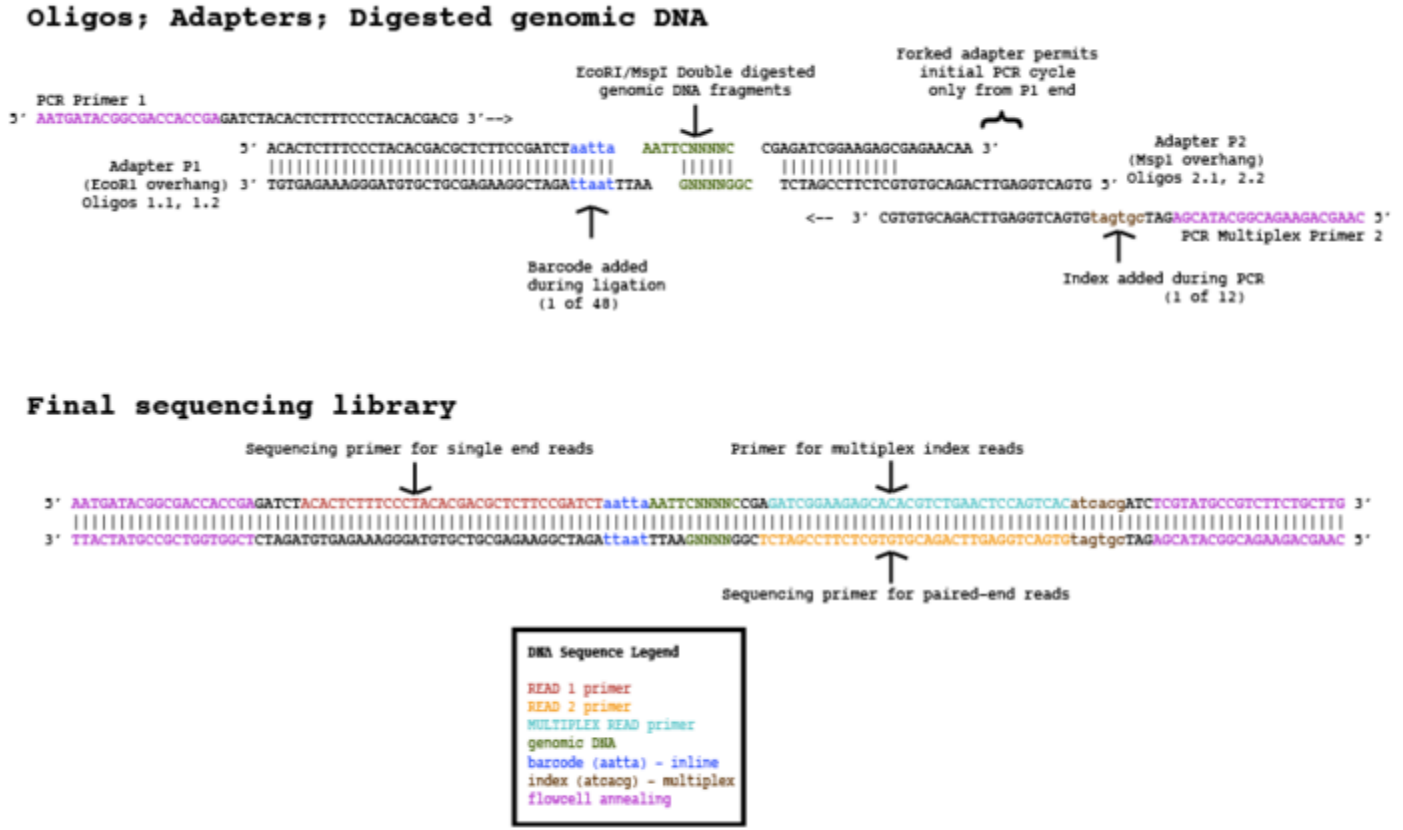
Harnessing the power of RADseq for ecological and evolutionary genomics
Nature Reviews Genetics, doi:10.1038/nrg.2015.28

Illumina library preparation: ddRAD



red regions will sequence because:

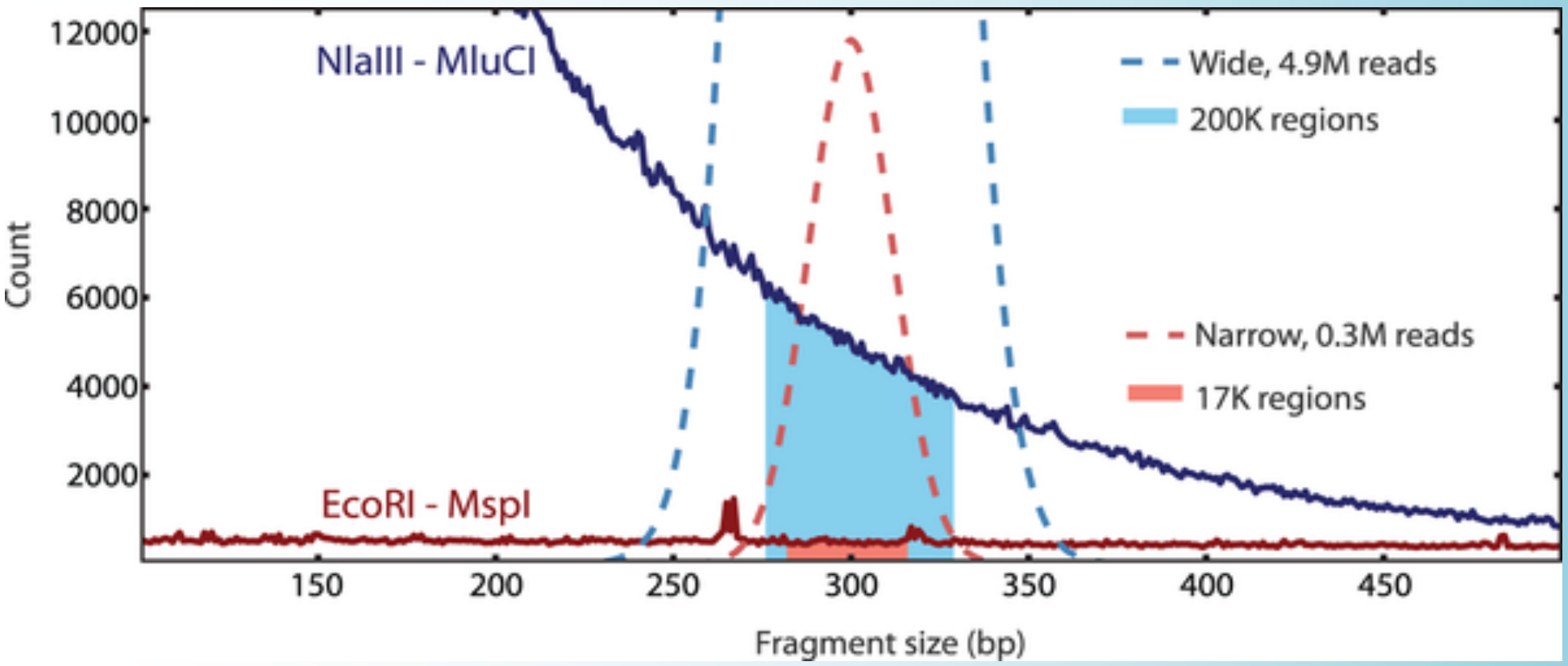
1) they are in the correct size selection range
2) they are flanked by one of each RE cut site



Selecting Enzymes and Size-Selection Range Considerations

- How many markers are needed/desired to answer your question?
  - REs with short recognition sites (4 or 6 bps) will cut more frequently than REs with longer recognition sites (8 bps)
- Is size of genome known?
- Is there a reference genome?
  - If genome is available, can use in silico digestion
    - SimRAD R package (DOI: 10.1111/1755-0998.12273)
  - If genome is not available, can test single and double RE digests on bioanalyzer
    - instructions in http://www.bit.ly/ddRAD

Figure 3. Double digest RAD sequencing provides flexibility in the number of homologous fragments recovered.



Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE 7(5): e37135. doi:10.1371/journal.pone.0037135
http://journals.plos.org/plosone/artide?id=info:doi/10.1371/journal.pone.0037135

Changing the restriction enzyme (RE) or size-selection regime modifies the fraction of genome recovered. Simulation 1 (blue lines, shading): the expected fragment size distribution for a RE digest with NlaIII and MluCI (CATG and AATT) in the Mus musculus genome is shown (solid blue line). "Broad" size selection (300 bp±50 bp) is modeled by a normal sampling distribution (mean = 300 bp, SD = 25 bp). Under this sampling distribution, 4,900,000 sequence reads (dashed blue line) are expected to cover ~119,000 regions at 7× or greater (blue area). Simulation 2 (red lines, shading): the expected fragment size distribution for a digest with EcoRI and MspI (GAATTC and CCGG) is shown (solid red line). "Narrow" size selection (300 bp±24 bp; see text) is modeled by a normal sampling distribution (mean = 300 bp, SD = 11 bp; see Analysis S1 Supporting Figure 1). Under this sampling distribution, an investment of 315,000 sequence reads (dashed red line) is sufficient to recover ~17,000 regions at 7× or greater (red area).

General RAD-Seq advice

- Resuspend/elute DNA when extracting into a no EDTA buffer (10mM Tris pH 8-8.5). EDTA may impact enzyme efficiency during digestion.
- Be aware of available adapters; selecting restriction enzymes that match adapters already available in your lab can save a huge amount of money.
- Consult with Pippin Prep operators to make sure the width and position of the region to cut is within the operational range of the instrument. A narrow range is not always possible and even when that can be programmed, real world results are often wider than expected. Typically 100bp is the smallest size range that can be cut.
- Prepare as many similar samples together as possible (considering adapter availability, input limitations, similar quality/quantity) for sequencing to have better uniformity. It is sometimes difficult to integrate a second set of samples into a project later on.
- As with gDNA and RNA libraries, do some tests cycles or qPCR to determine the minimum number of cycles required to have sufficient material for sequencing.

Size-selection with Pippin Prep

Talk with the FGL in advance of submitting your sample to get advice on what the best settings would be:

- are you only trying to eliminate adapter dimer?
- are you trying to maintain as much material as possible in your size range?
  - tighter sizing windows are more difficult to accurately attain and may result in lower yields
  - from past experience Karen (lundyk@berkeley.edu) may know the best machine parameters to input to get you the real-world result you want
- or is it more important that you prioritize collecting sample *only* at a particular size?
- are you submitting these samples for sequencing immediately after size-selection?
  - If not, the FGL can return the sample without bead cleaning which may increase yields

RAD-Seq advice from the experts at Cal Academy

- Genotype more samples than needed with the expectation of having to drop those with low coverage/low SNP count. If you leave these samples in final dataset, you will be more likely to reduce the number of SNPs shared across libraries.
- We have found we get significantly better sequencing results when we standardize the amount of DNA going into the adapter ligation step. We usually try for 100 ng per sample (*Note: as quantified after digestion and bead clean-up. They start digestion with 300-400ng*).
- Standardize the number of samples per size-selection pool. In cases where we have had one pool with less individuals (or less DNA), despite our efforts to put proportionally more of that sample in the final sequencing pool, our sequencing results never come out with even coverage. The bottom line is, as much as you can, try to put the same number of individuals in a pool and keep the amount of DNA going into ligation the same for all pools.
- DNA quality of samples is important, especially among pooled samples. **Do not pool samples for size-selection that have dramatically different DNA quality** (i.e HMW vs. fragmented). Closely scrutinize DNA quality of samples prior to starting library prep: run on agarose gel, Nanodrop and Qubit.
- When DNA of similar quality and quantity are pooled together for size-selection, their respective concentrations can be brought to similar levels during enrichment PCR by altering the number of cycles. When pools contain both low and high quality samples, individual samples don't often PCR the same which results in some samples having much less coverage than others
- Keep number of individuals/lane conservatively low so as to ensure high coverage. Or be prepared to have to sequence an additional lane if coverage is too low.

The bottom line is, as much as you can, try to keep the amount of DNA going into ligation the same for all individuals, put the same number of individuals in a size-selection pool, and only pool individuals of the same DNA quality

General Library Preparation Advice: all methods

- Read the protocol carefully, many times through and consult with someone who has used the same protocol before for more specific advice
- Do a test run of the whole protocol with samples that are unimportant or repeatable
- Take your time: all future success relies on this stage
- Plan your indexes carefully
- More is not always better. For PCR-required methods, use as few cycles as possible: enough to fully incorporate the index, to allow the library to be easily quantified, and to have material available as a back-up but no more.
  - This depends on starting material, but usually 8-10 cycles gDNA (two-step), 10-15 cycles RNA.
- Be conservative to avoid overamplification in pooled samples (ex. post-capture enrichment, ddRAD post-size-selection) which can lead to an overabundance of PCR duplicates in sequence data and/or to index swapping and chimeric molecules
- For libraries that will be used for captures, resuspend the amplified reactions in water rather than buffer since they will be concentrated before hybridization

General Library Preparation Advice: all methods—SPRI beads

In many respects, library preparations are bead clean-ups

- Represent 50% (or more) of researcher hands-on time
  - Robotics available in many Berkeley core labs for very large projects
- Used to rid reactions of:
  - Enzymes
  - Oligos, adapters, and dNTPs
  - Undesired sizes of DNA inserts (too large or too small)
- SPRI beads work based on volume ratios rather than the total mass of DNA added.
  - Counterintuitive: 100µL of DNA at 10 ng/µL and 100µL of DNA at 1 ng/µL will get cleaned with the same volume of beads

General Library Preparation Advice: all methods—SPRI bead handling

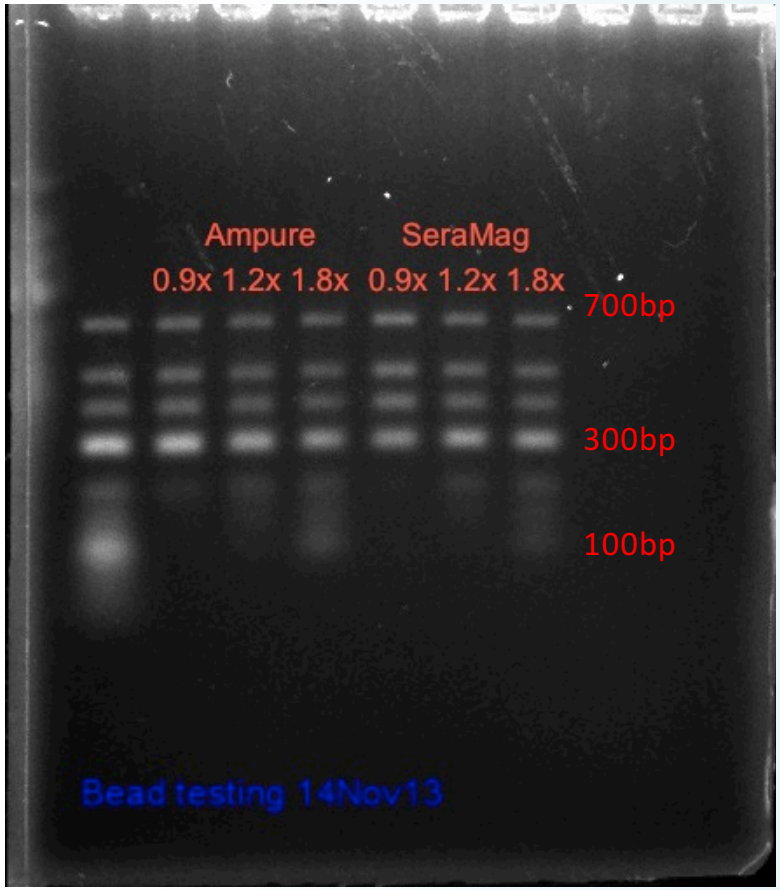
- Always test bead formulations at your protocol's required ratios to ensure that they behave as predicted
- Warm beads to room temperature before any clean-ups
- Always make fresh dilutions of 80% ethanol each day for bead clean-ups
- Increase incubation times when using low ratios of beads. (If using lab-made beads, a special formula may be needed at very low ratios.)
- The most critical timing step is drying ethanol off the beads after the washes. It is very important not to over-dry or yields will dramatically decrease
  - I manually remove all residual alcohol in the well with a small pipette tip and/or blot with an autoclaved toothpick. Then I resuspend immediately while the beads are still shiny
  - If beads do not easily go back into solution, that is a sign that they dried for too long. Next time, shorten the drying time
  - When cleaning many rows of samples at once I may only dry 16 or 24 samples at a time to avoid the first ones over-drying by the time I am ready to elute them
- Elute in 10mM Tris (pH 8-8.5) or water at a similar pH

How do these magical SPRI beads work?

- There are always enough beads to bind all the DNA in a sample no matter how much/how little SPRI solution is added (within reason)
- Changing the ratio changes the percentage of other factors in the solution like PEG and NaCl which control what length of DNA can bind to the beads
- The lower the ratio of solution added, the higher the size cut-off:
  - http://www.keatslab.org/blog/pcrpurificationampureandsimple
  - http://core-genomics.blogspot.com/2012/04/how-do-spri-beads-work.html
  - http://blog.genohub.com/peg-size-selection-and-precipitation-of-dna-libraries-how-ampure-or-spriselect-works/

Lab-made SPRI beads

Rohland, Nadin, and David Reich. "Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture." *Genome research* 22.5 (2012): 939-946. DOI: [10.1101/gr.128124.111](10.1101/gr.128124.111)



Work equivalent to Ampure and other commercial beads at typical medium to high ratios (> 0.9x)

Cost a fraction of commercial SPRI beads: ~$400 for 750mL

Always test to insure they behave properly at your intended ratios
Caution: test with Fermentas GeneRuler ladders. Others reported to not behave properly

SPRI beads: double-sided selection

Double-sided selection is a good method for refining imprecise fragmentation, but a lot of material can be lost (even in the target range)

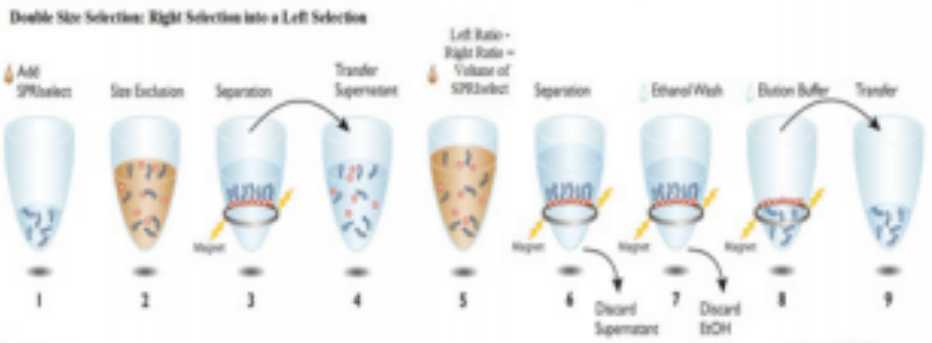
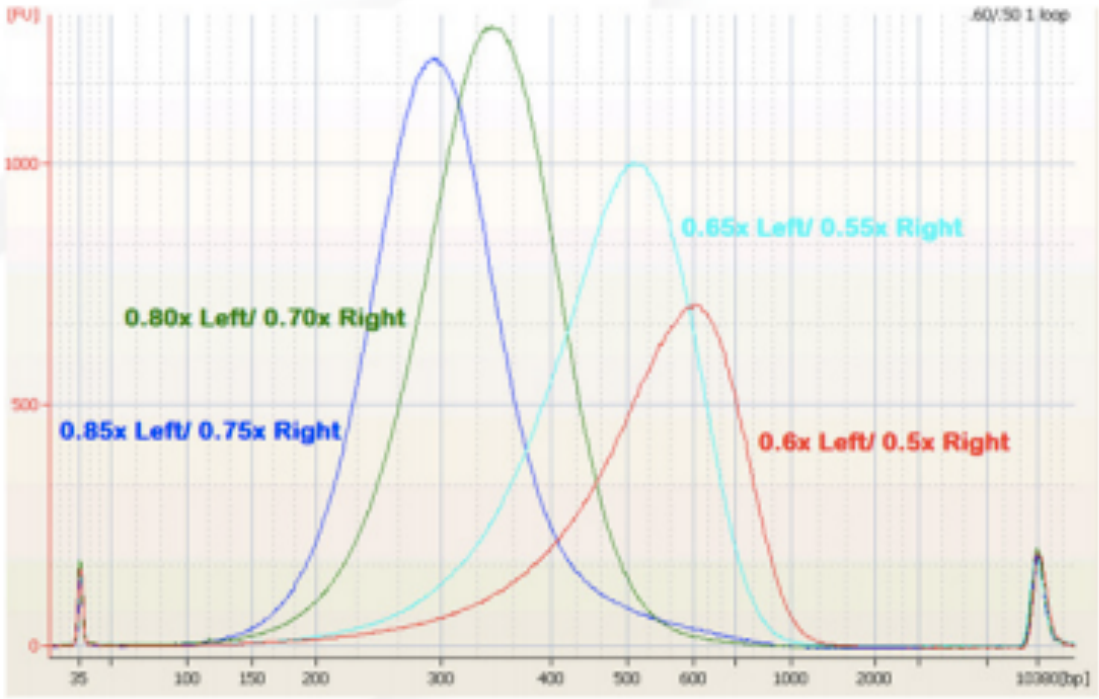


Figure 5 (Left): Process overview for Double Size Selection using the Right into Left method.

Figure 6 (Right): The sheared E. coli DNA has been processed using four different ratio sets for Double Size selection.



Requires commercial (Ampure, Kapa, etc) or special low-ratio formula SPRI beads

	RNA Seq/Transcriptome	RAD-Seq	Amplicon sequencing
EGL sample prep costs	$30-45 for poly-A selection; $100+ for rRNA depletion	very cheap after initial oligo investment for large numbers of samples (~$10-15)	very cheap after initial oligo investment for large numbers of samples (~$5-10 for reagents)
museum/historical/degraded samples?	no	no	no
special equipment?	none required (homogenizer/bead beater useful)	Pippin prep (available at the Functional Genomics lab in LSA)	none for metagenomic studies. Emulsion PCR or fluidics required for large, multilocus projects
reference genome/prior genomic information required?	useful but not required for initial data	useful to have a closely related reference genome for in silica RE tests, but not essential	no reference genome needed, but PCR primers for the areas of interest must be available.
major start-up costs	Y-shaped adapters	bar-coded adapters	adaptors and PCR primers
main benefits	RNA Seq: differential gene expression and transcript variants. Transcriptome: genomic reference and comparative exomic data	generates unbiased, genome-wide set of markers for SNP detection; degree of genome reduction is manipulatable based on RE selected.	metagentic studies (multiple organisms per amplicon)
main drawbacks	expense of library prep; only exons sequenced; RNA not always available in tissue samples; highly expressed genes dominate seq data and can make rare transcript sequences and isoforms difficult to identify	cannot be targeted to specific regions of the genome; homology may difficult to obtain between more distantly related taxa due to mutations at RE cleavage sites; problem of uniformity of results	amplicon length determined by current Illumina PE max (300 on MiSeq); very labor intensive / costly beyond a few loci

	array-based capture	in-solution capture (Nimblegen)	in-solution capture (MyBaits)
EGL sample prep costs	$15-25	$15-25	$15-25
museum/historical/degraded samples?	yes	yes	yes
special equipment?	hybridization oven	Cycler with adjustable lid	Cycler with adjustable lid
reference genome/prior genomic information required?	reference genome or transcriptome required from a closely related species	reference genome or transcriptome required—may be from a more distantly related species (ex. using Xenopus exons to capture other frogs)	reference genome or transcriptome required for custom kits; none for UCE or mtDNA capture
start-up costs	array: $750 capture reagents: $150	kit: $8000 (5 reactions) or $13,000 (14-15 reactions) capture reagents: $120/rxn	kit: $700 (8 UCE rxns) kit: $4000 (16 custom rxns) and up capture reagents: $120/rxn
main benefits	low per-capture cost for a lot of data (1 million custom probes); pilot projects (proof of concept); great for population genetic studies where no sample is very distant from the probe seq	large-scale capture projects; very high mapping efficiencies; 2.1 million custom probes allow very high target sizes and capture multiplexing	lowest cost in-solution capture method; very small buy-in for pre-designed kits; more capture reactions available than Nimblegen; ideal for phylogenetic studies
main drawbacks	array has technical challenges; lower mapping efficiencies than in-solution; needs reference genome or transcriptome; probes should be <5% divergent from targets; need for species-specific cot-1; arrays being phased out by vendors (still available from Agilent)	high kit prices require very large initial investment*; needs reference genome or transcriptome for designing probes *New EZ Choice program gives the Nimblegen sales reps more flexibility with pricing for new customers. It can hurt to communicate your budget and ask for a discount	Far fewer probes than Nimblegen so multiplexing is reduced (~10-20 libraries max); Capture reagent costs add up; reference genome or transcriptome needed if not using UCEs (and not all organisms have a probe set developed)

Library quality control

Prior to sequencing, the following quality control checks are recommended or required:

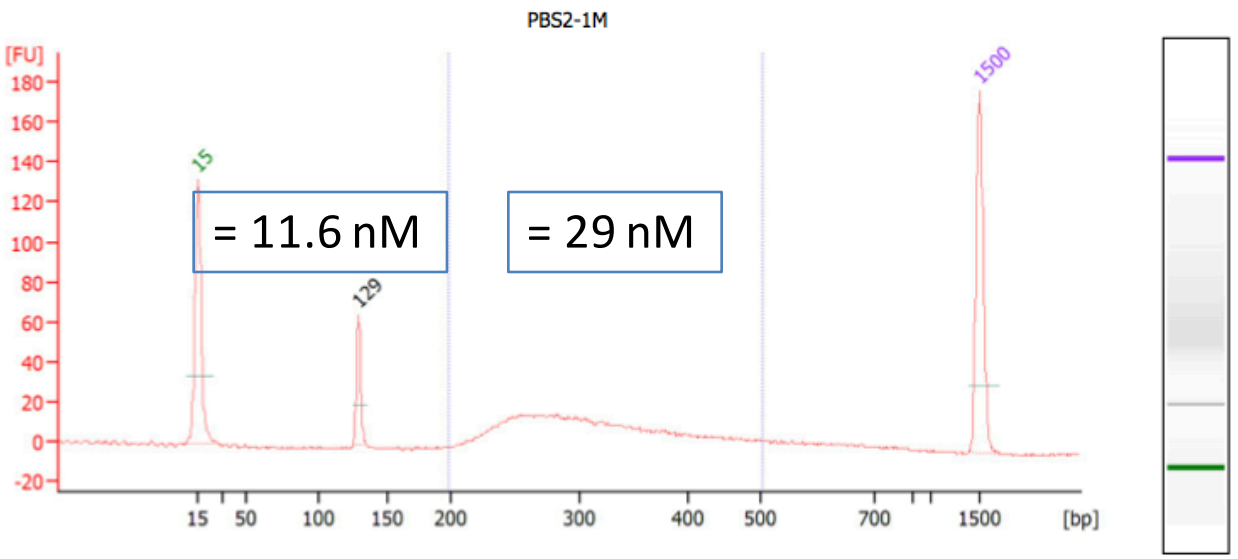
1) Bioanalyzer provides length information (required with GSL submission)
2) Qubit provides concentration information (required with GSL submission)
3) qPCR with standards provides molarity information (usually performed by GSL)

Bioanalyzer and qubit are available at the FGL and EGL or can be run after sample submission by the GSL. I recommend assessment before submission so that any problems can be fixed in advance (samples concentrated; adapter dimer removed, "dud" samples removed/replaced.) But the GSL will notify you if a submission cannot be sequenced.

qPCR is usually run by GSL (one free qPCR per lane, others $12) but may be done in the EGL if a researcher needs to pool many samples

Library quality control: bioanalyzer cDNA library

Electropherogram Summary Continued ...



Overall Results for sample 4 : PBS2-1M

Number of peaks found:	1	Area 1:	290.1

Peak table for sample 4 : PBS2-1M

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	129	0.99	11.6	
3	1,500	2.10	2.1	Upper Marker

Region table for sample 4 : PBS2-1M

From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
199	502	290.1	78	320	21.7	6.17	

Illumina adapter dimer ~130bp single index, ~140bp dual index

Library quality control: adapter dimer

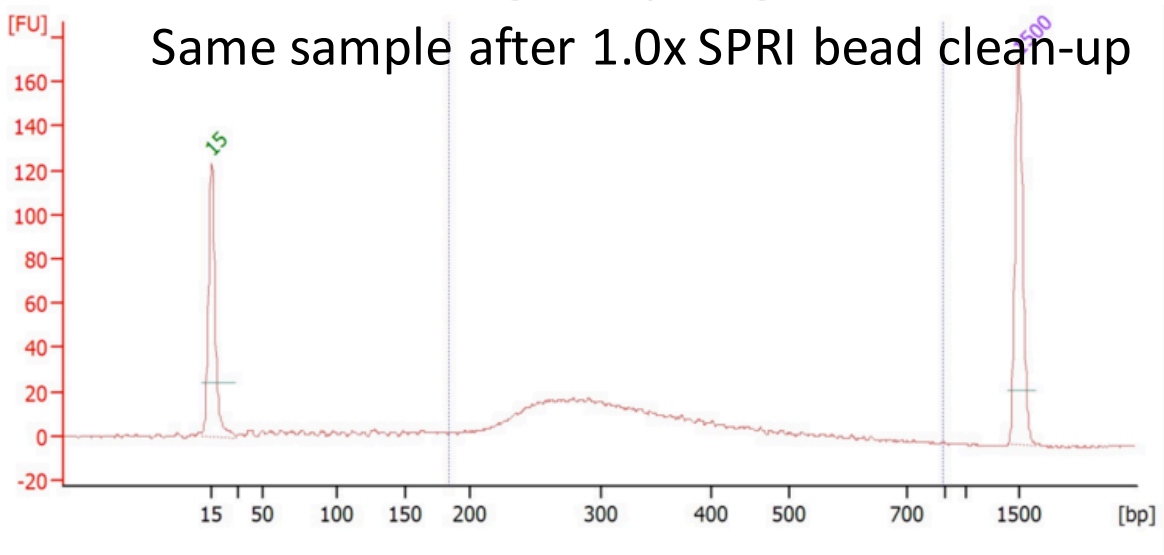
- Adapter dimer will steal a disproportionate amount of reads away on the Illumina flowcell.
  - Since they are small they have a high molarity (disproportionately more fragments) and smaller fragments attach better to the flowcell
- It will also make it more difficult to accurately quantify your true library with qPCR and will lead to poorer quality data if > 5% of total reads are from adapter dimer
- Miniscule amounts of dimer may be okay, especially if your actual libraries are relatively small
- Check with the GSL team before submitting libraries for sequencing to see if they recommend you do an additional clean-up
- Shana McDevitt says: “Generally if I can see it, you should remove it.”
  - Exception if the yield of the library after additional cleaning can’t be preserved above the minimum requirement (10µL of 3nM)
- Usually an additional bead clean-up at a low ratio ≤ 1x is sufficient
- Or size-selection with Pippin Prep at the FGL,

Library quality control: bioanalyzer cDNA library

Electropherogram Summary Continued ...

KS001J [dimer cleanup 1x beads]

Same sample after 1.0x SPRI bead clean-up



Overall Results for sample 3 : KS001J

Number of peaks found:	0	Area 1:	365.5

Peak table for sample 3 : KS001J

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	1,500	2.10	2.1	Upper Marker

Region table for sample 3 : KS001J

From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
184	844	365.5	86	335	30.1	7.86	

GSL testing of adapter contamination on the HiSeq 4000

100SR Lane	Read Count	% Adaptor Dimer in Reads	Error Rate (aligned)
400 broad (.8%)	372M	6%	0.15
400 broad (2.4%)	380M	8%	0.14
400 broad (9.6%)	357M	25%	0.17
400 sharp (1.3%)	362M	8%	0.18
400 sharp (9.3%)	344M	25%	0.15
600 sharp (4.1%)	360M	37%	0.20
600 sharp (8.5%)	348M	55%	0.28

Table courtesy of Shana McDevitt

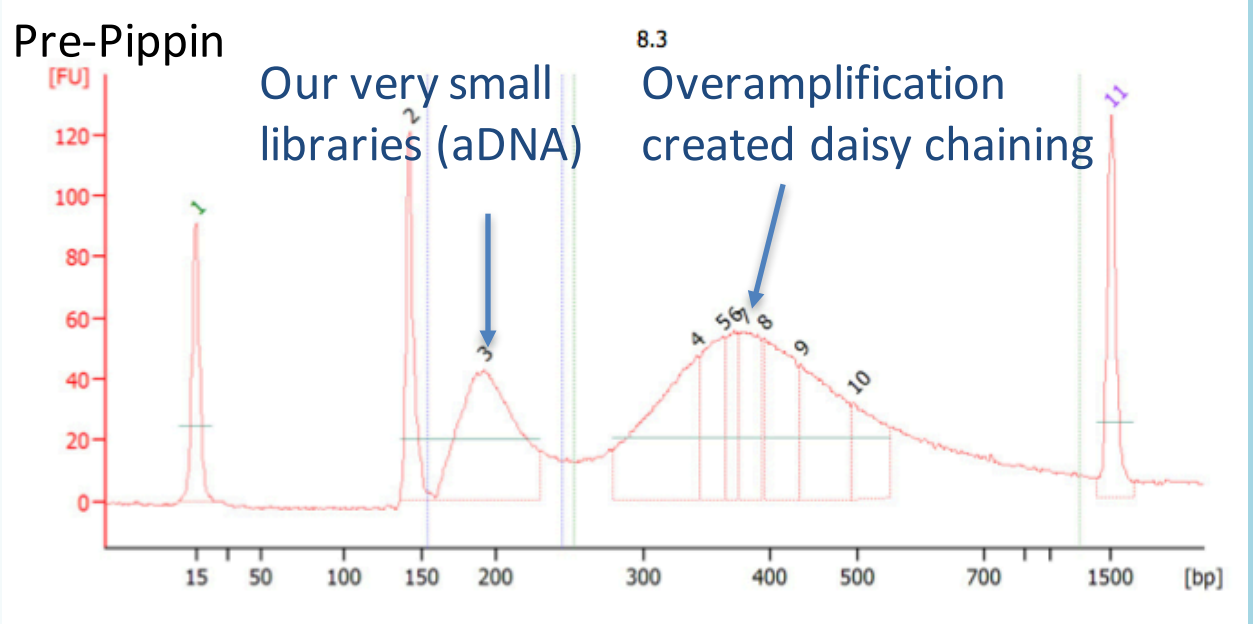
- RECOMMENDATION: sequence libraries with no more than 0.5% dimer visible
- As little as 0.8% dimer seen in bioanalyzer traces can lead to > 6% dimer in reads
- The larger the average library size, the bigger the impact of bioanalyzer dimer % on read %
- The more dimer in reads, the worse the error rate

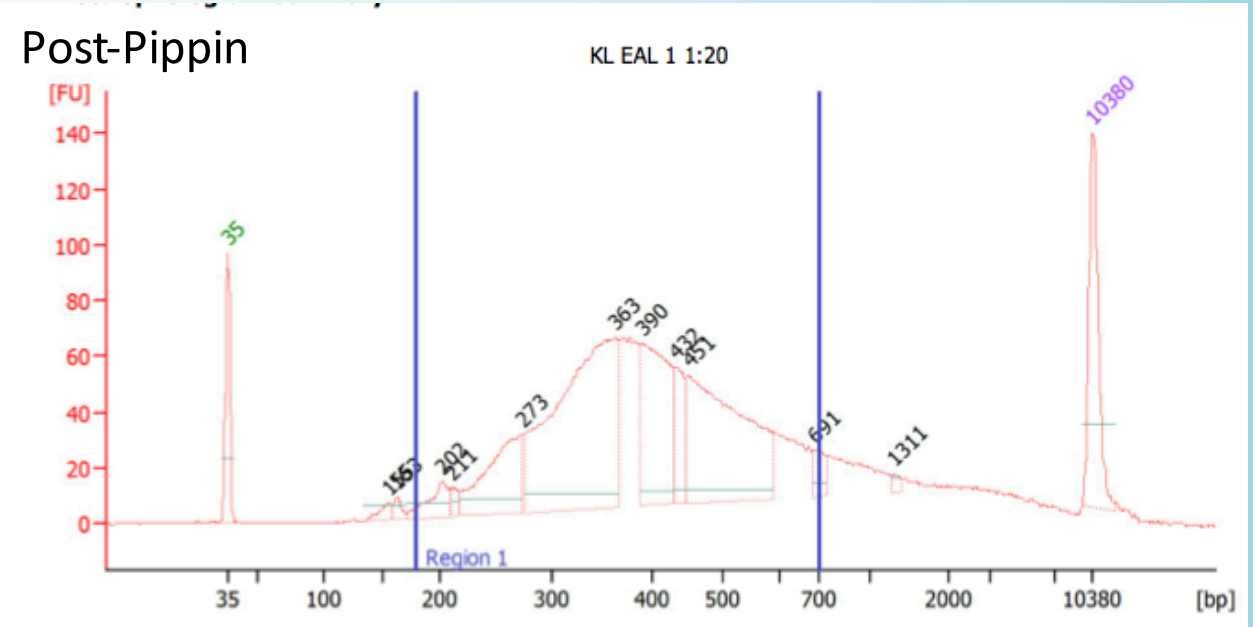
Pippin Prep for adapter eradication

Note: Pippin Prep can cause library fragments to associate in strange and deceptive ways.

Although the excision worked very well to get rid of the majority of the adapter dimer, the use of the bioanalyzer for sizing is no longer accurate

Important that I communicated that the true average library size should be about 200bp (rather than 400bp) so that the GSL staff could make the proper molarity calculations after qPCR





Library QC: Index switching on flowcell

- Small amounts of free indexing oligo or adapter may cause a phenomenon called “index switching” or “index hopping”on the Illumina flowcell
  - Reads from one sample are assigned the index corresponding with a different sample (!!!)
- Proposed mechanism in Sinha, et al. pre-print: https://doi.org/10.1101/125724
- Illumina documentation:
  - https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf
  - https://www.illumina.com/science/education/minimizing-index-hopping.html
- Depending on the prep library method and cleanliness of the library, rates vary:
  - Sinha et al. found rates of 5-10% with their RNAseq protocol
  - With EGL protocols we see far less:
    - Single-index capture libraries <0.2% (average around ~0.15%)
    - Single-index RNAseq: 0.2-1.5%
    - Single-index RAD: 0.2-2% (average is less than 1%)
- If any free adapter or primer can be seen in the bioanalyzer trace, do an additional low-ratio SPRI clean-up (0.8x) or excise with the FGL’s Pippin Prep: “*Ugly things can happen when sequencing really ugly libraries*”

Library QC: Index switching on flowcell

- Even a small amount of index switching can be highly detrimental to low coverage projects
  - There is an especially high risk of occurrence in PCR-free WGS because there are so few steps following adapter ligation
- Recommended: always uniquely dual-index libraries when resources are available to do so.
  - GSL sells 384 uniquely dual-indexed adapters for use with PCR-free DNA kits and many RNAseq kits
  - "swapped" data will be bioinformatically removed when demultiplexing
- If it is not possible to give each library a unique dual index, pick dual indexing options that will at least minimize the impact
  - Ex. if I have access to 96 P7 indexing oligos and 8 P5, for a set of 192 libraries, I can organize the indexing oligos so that index switches can at least be identified 75% of the time.

Library quality control: molarity estimates

- Sequencing facilities usually ask for library submissions at ≥ 10µL of 10nM
- The molarity of a sample can be estimated using the concentration given by the qubit and the average length given by the bioanalyzer.
  - http://www.promega.com/a/apps/biomath/index.html?calc=ugpmols
  - Determine the number of pmol represented by 1 µL of your sample pool
  - This number is equivalent to µM; multiply by 1000 to get nM
- This molarity will be an underestimate since the qubit can't distinguish between full-length sequence-ready libraries and other double-stranded DNA
- But most 10nM by qubit libraries will be ≥ 3nM when qPCRed

dsDNA:

Formula

$$\mu g\ DNA \times \frac{pmol}{660pg} \times \frac{10^6 pg}{1\mu g} \times \frac{1}{N} = pmol\ DNA$$

N is the number of nucleotides and $\frac{660pg}{pmol}$ is the average molecular weight of a nucleotide pair.

Library quality control: molarity estimates

Shortcut between concentration and molarity estimate once average library length is known:

$$nM = \frac{\text{concentration (ng/µL)} * 10^6}{660 * \text{avg length}}$$

So, a dsDNA library of 10 ng/µL with an average length of 500bp (~360bp insert and ~140bp dual-index adapter) is about 30 nM

Library quality control: qPCR with known standards

- The **best** way to know the molarity of your library is to do qPCR against known standards
- Still need bioanalyzer data for average length
- qPCR will amplify only sequence-able fragments and will include any full-length single-stranded library fragments
- Part of standard Berkeley GSL library quality control: one qPCR assessment free with every lane submission.
  - If multiple libraries are to be pooled in a lane, the GSL can assess multiple samples for $12 each and then pool in the desired proportions

Library submission to GSL

- Before making aliquot tubes for the GSL, place samples on a magnet in case of residual bead carry-over
- Use siliconized or low-binding microcentrifuge tubes for sample storage
- GSL Submission guidelines (read these before starting your library preps):
  - http://qb3.berkeley.edu/gsl/sequencing-samples/
- Also, be prepared to acknowledge the GSL in your publications: http://qb3.berkeley.edu/gsl/faqs/ (check with other facilities directly about the best way to acknowledge their contributions to your research.)

How many samples can be pooled together in a single lane? DNA

Warning: this is just an example of the thought process. Use the numbers you calculate to start talking to other people and compare your theoretical results with real-world examples from similar projects

Start with the projected amount of data: (HiSeq 4000 PE150, 350M reads)

105,000,000,000 bases

Multiply by expected *conservative* on-target percentage for unique fragments (say 25% for in-solution capture):

26,250,000,000 bases

Divide by the target length (3.125 Mb):

8,400

Divide by the amount of coverage you desire (be conservative: this will give only the average expected coverage but it is never even across all target) (50x):

168 samples for multiplexing

How many samples can be pooled together in a single lane? DNA

What if all samples can't fit on the same lane?

In the previous example we calculated a conservative estimate of 168 samples/lane but you have 384 samples:

1) When planning your indexing strategy, give each sample a unique index (ideally, unique dual index) so that all finished libraries can be pooled together

2) Consult with the GSL about pooling all of your samples together.

3) Creating a pool of all the samples in a project and spreading them out over multiple lanes eliminates the impact of lane effects on your data

4) In this situation, if time is not of the essence, run 2 lanes of data collection, then analyze to see if that gives enough coverage for your statistical needs

--if you run 3 lanes of data but only needed 2, that money cannot be recovered

--but if you undersequence and all samples are pooled together, adding an additional lane is easy

How many samples can be pooled together in a single lane? RNA

Same warning as previous example

Often for RNA-Seq projects researchers think in the number of reads per samples

What questions are you trying to answer?

- Are you interested in just the sequences of the 50% most highly expressed transcripts (as low as 1M reads/sample: doi:10.1016/j.gene.2014.12.013)?
- Are you interested in comparing differential expression (~30M reads/sample)
- Do you want to discover novel elements or perform more precise quantification, especially of lowly expressed transcripts (100-200M PE reads/sample or even more)

Recommended to always pool all samples together (https://doi.org/10.1534/genetics.110.114983) then spread the pool out over multiple lanes in order to negate to influence of lane effects. Pre-plan to have sufficient indexes available.

https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

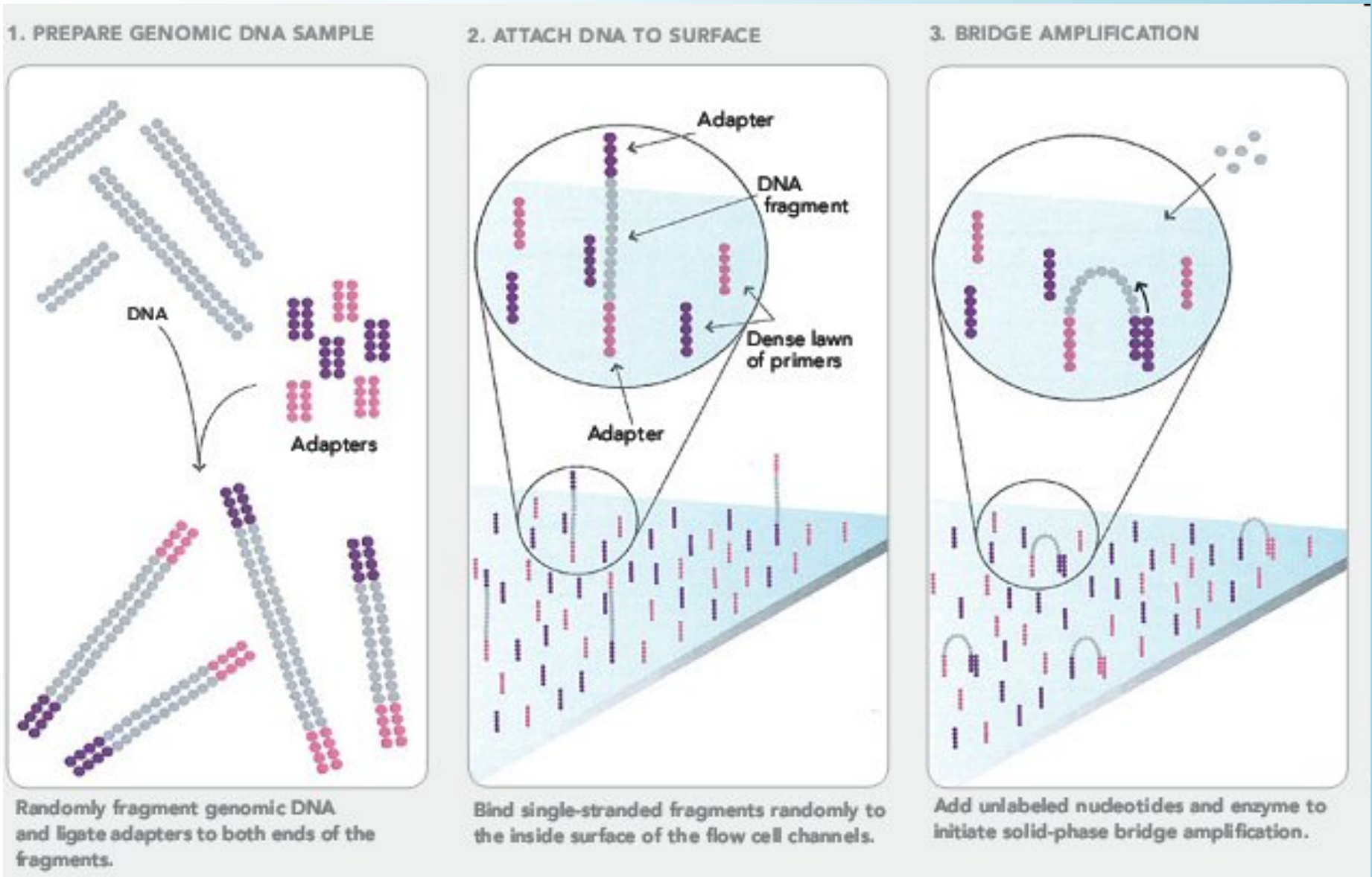
https://www.msi.umn.edu/sites/default/files/RNA-Seq%20Lecture_2016.pdf

HiSeq 4000 sequencing

- Reliably 350,000,000 clusters passing filter each lane (for PE150 = 105 gigabases of data if the two reads don't overlap)
- Limited to PE150 as the longest read
  - MiSeq = PE300 and HiSeq 2500 = PE250
- Not recommended for amplicons
- RAD-Seq okay with a ~10% PhiX spike-in
  - PhiX contributes diversity at restriction enzyme cut sites
  - (All Illumina libraries/platforms are run with at least 1% PhiX as a control)
- Poor choice for broadly distributed or bimodal insert sizes (unless the user accepts a sequencing bias towards small molecules)
- Submission must be > 3nM, as measured via qPCR with < 0.5% adapter dimer

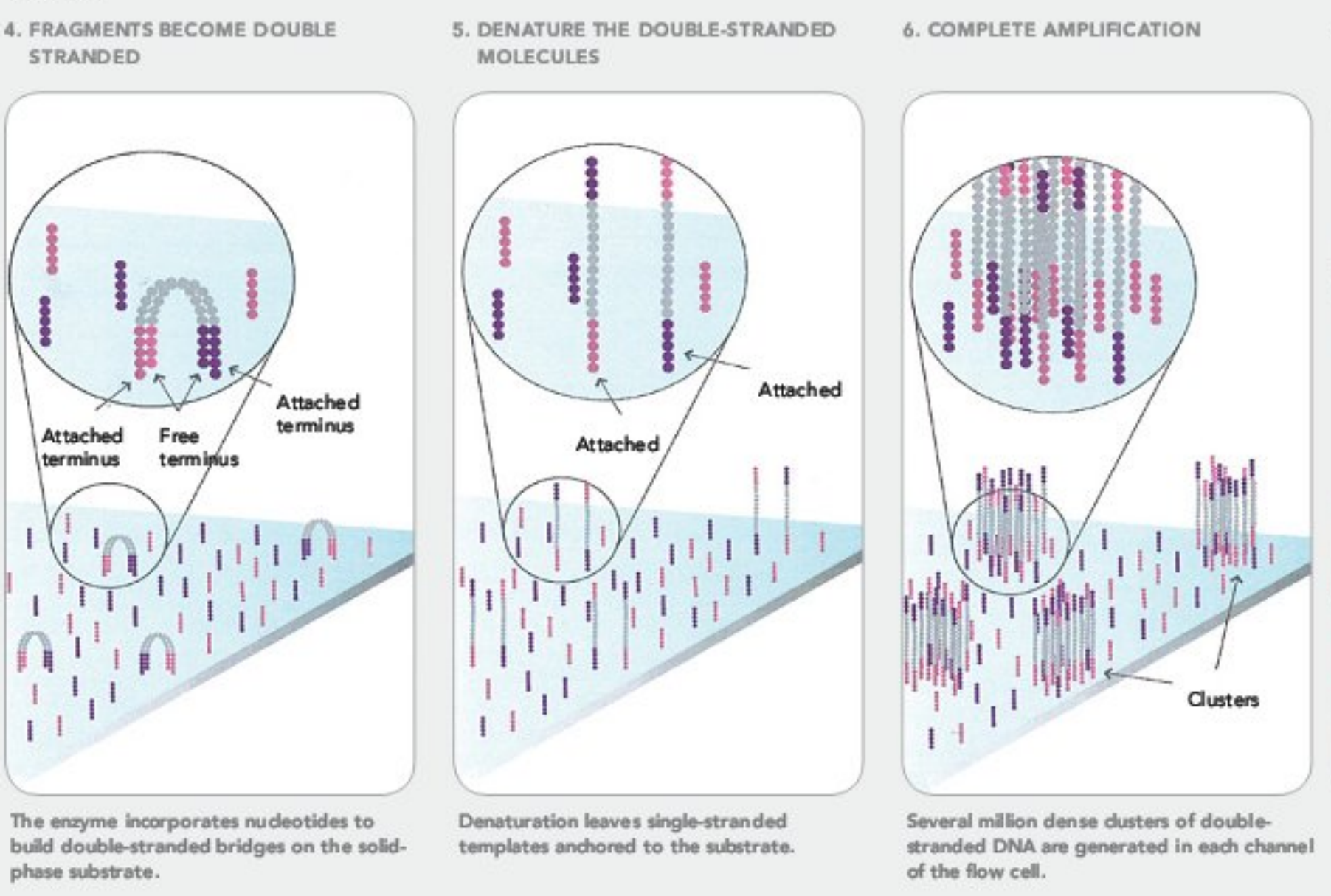
Illumina sequencing: how does it work 1

https://youtu.be/fCd6B5HRaZ8



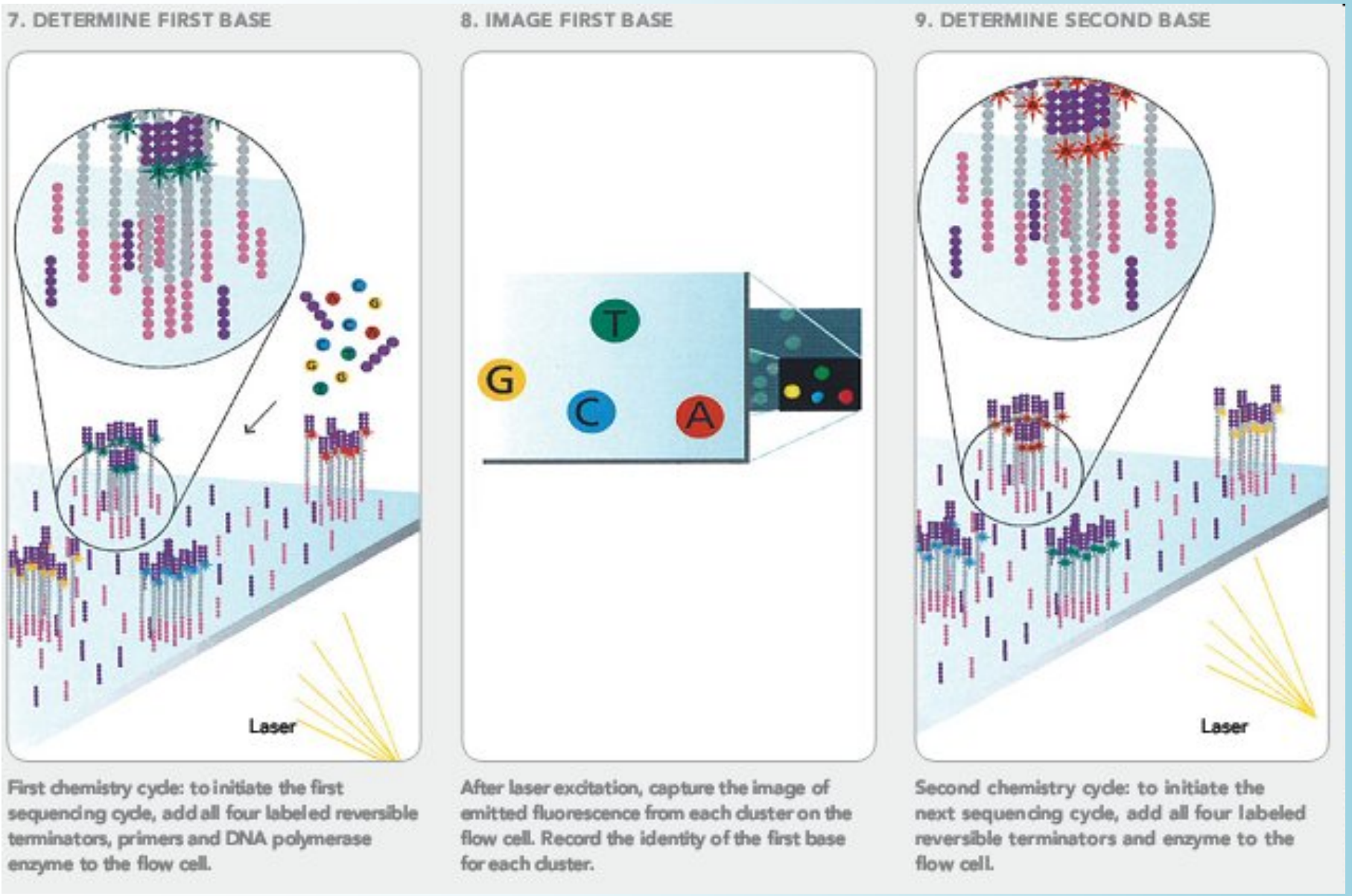
Illumina sequencing: how does it work 2

https://youtu.be/fCd6B5HRaZ8



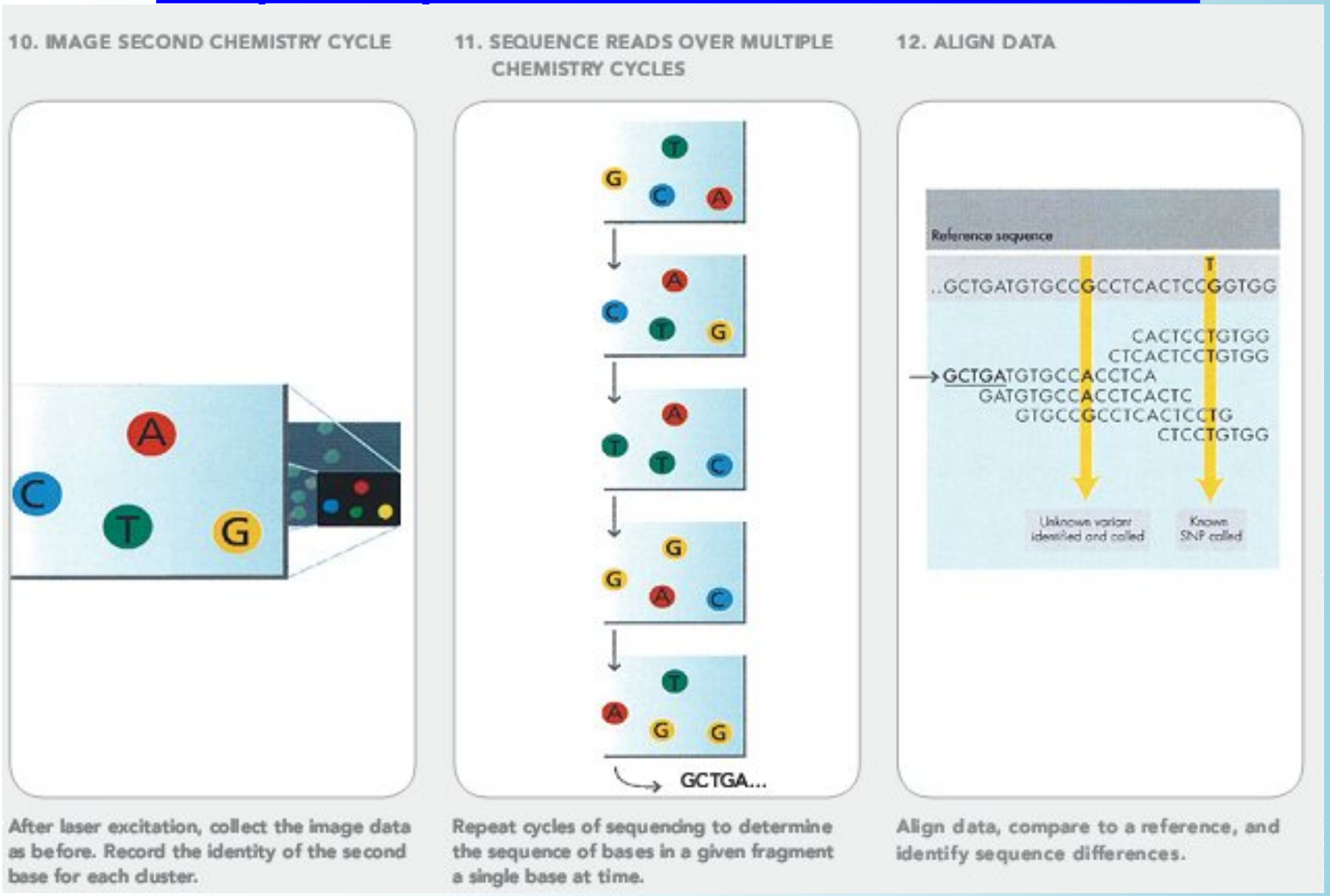
Illumina sequencing: how does it work 3

https://youtu.be/fCd6B5HRaZ8



Illumina sequencing: how does it work 4

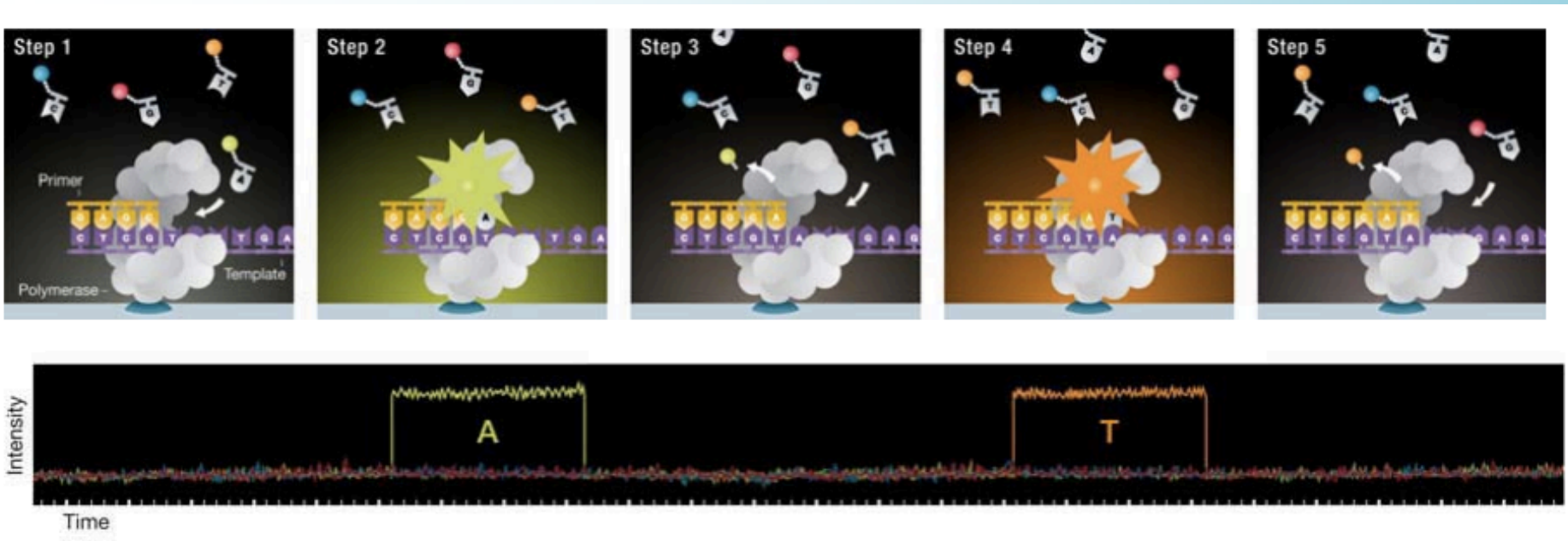
https://youtu.be/fCd6B5HRaZ8



PacBio Sequencing

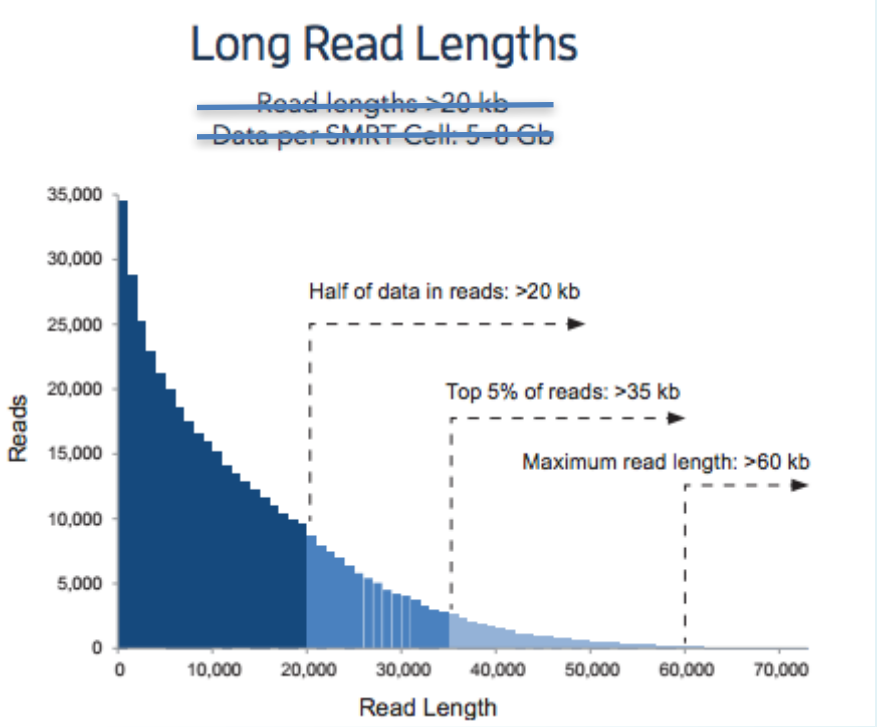
- Long-read sequencer available at Berkeley GSL (PacBio Sequel)
  - Whole genome sequencing
  - Iso-Seq (whole transcript sequencing)
  - Long amplicons
- For most projects, only a single to a few libraries are needed.
  - Long-insert preps are expensive ($450); best to send to a core facility with experience than to DIY
- But if you have high-quality and quantity of input and need long reads (1000's of bp), there is currently no more reliable platform
- *GSL seminar to come soon*

PacBio Sequencing:
Single Molecule, Real-Time (SMRT) Sequencing



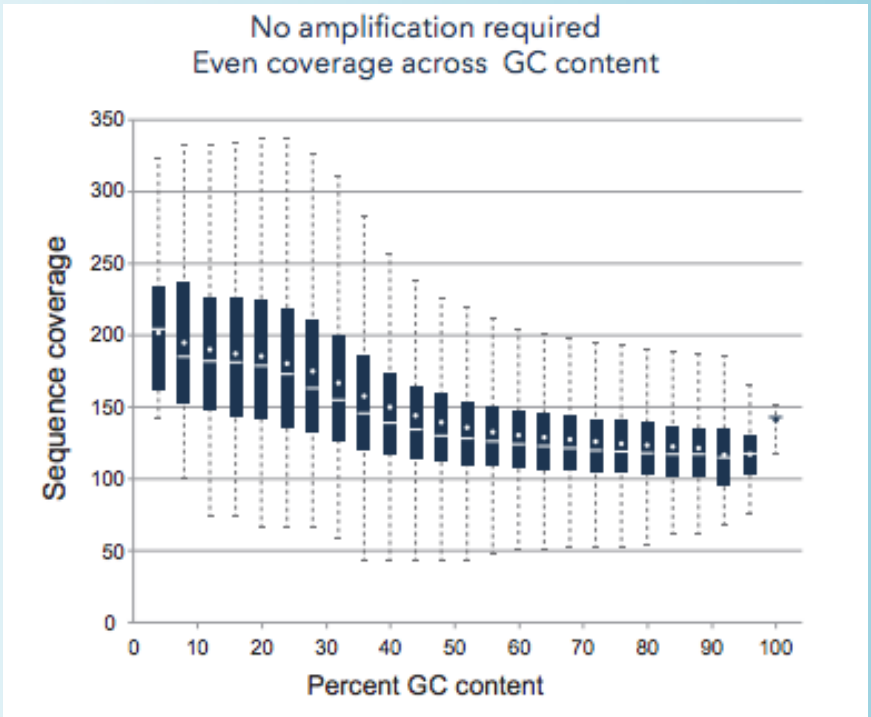
Step 1: Fluorescent phospholinked labeled nucleotides are introduced into the ZMW.
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing a bright flash of light.
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.
Step 4-5: The process repeats.

PacBio Sequencing: Advantages

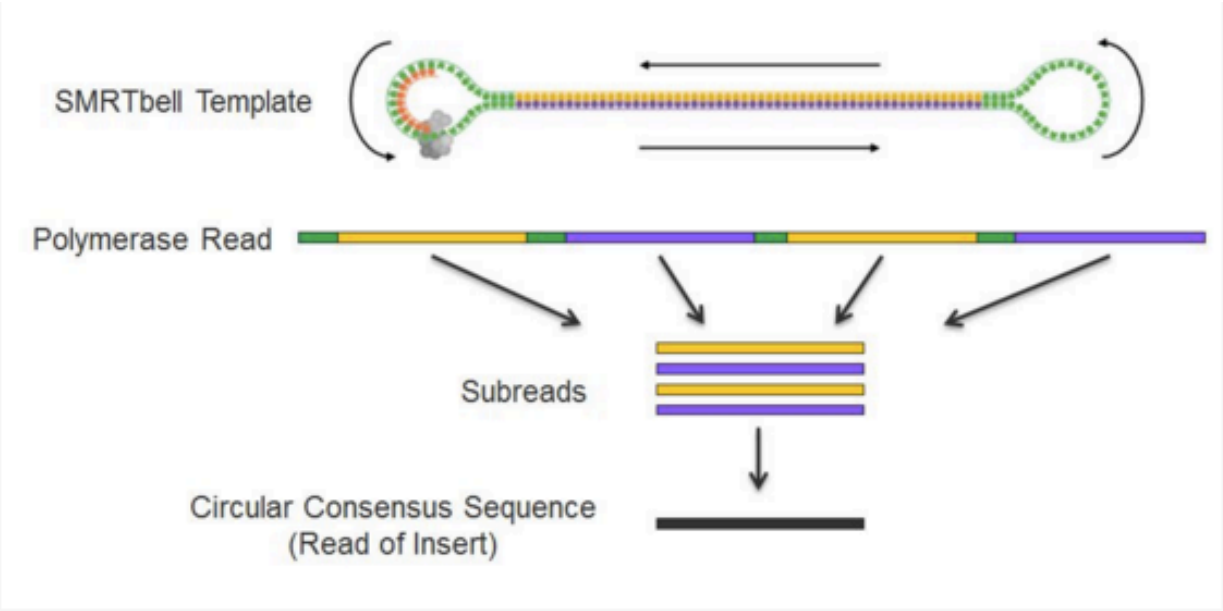


1) Very long read lengths!
Easily 10kb, up to 35kb

2) Single molecule sequencing:
No amplification biases

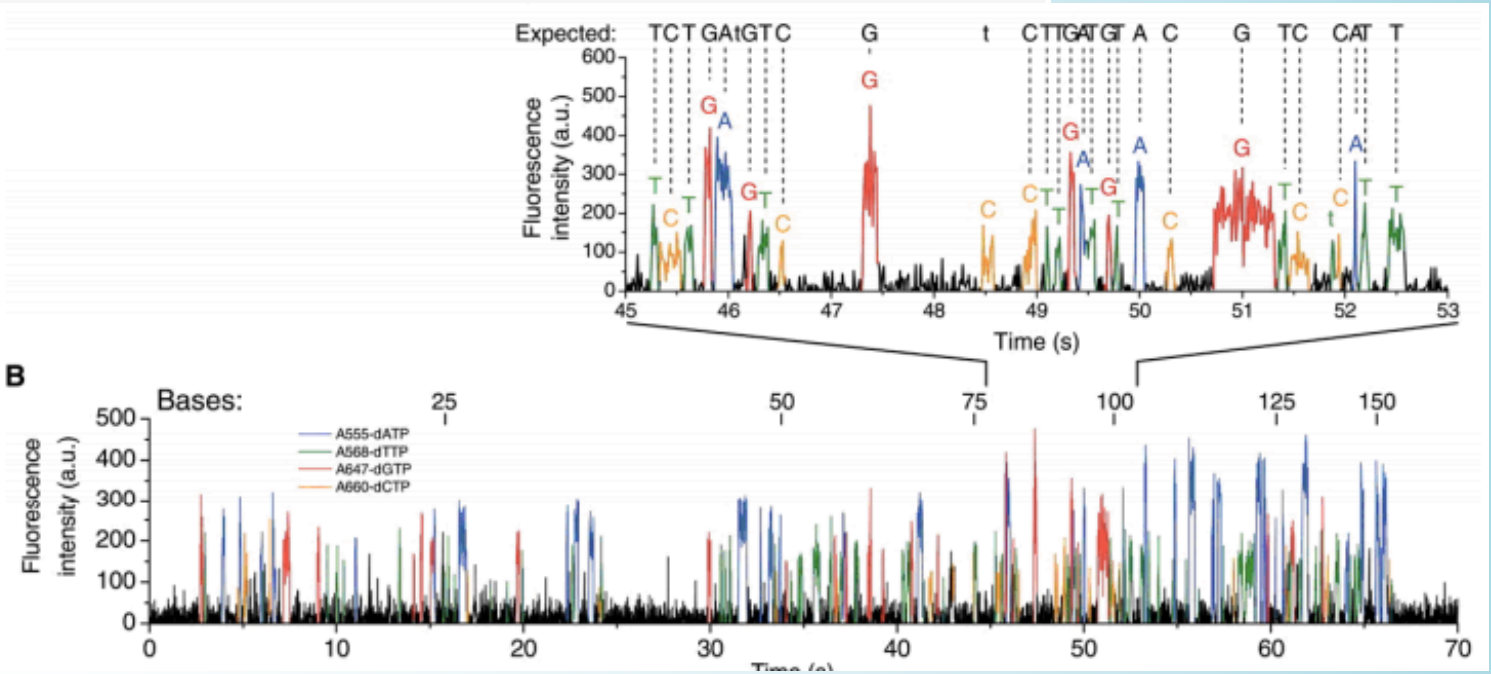


3) Has fewer difficulties with repetitive elements, high-GC content, homopolymers, and other challenging genomic regions



The circular nature of the SMRTbell DNA template allows polymerase to sequence the same DNA molecule multiple times with multiple passes. This produces high intra-molecular consensus accuracy.

Pac Bio Technology



https://www.youtube.com/watch?v=WMZmG00uhwU

PacBio Sequencing: Disadvantages

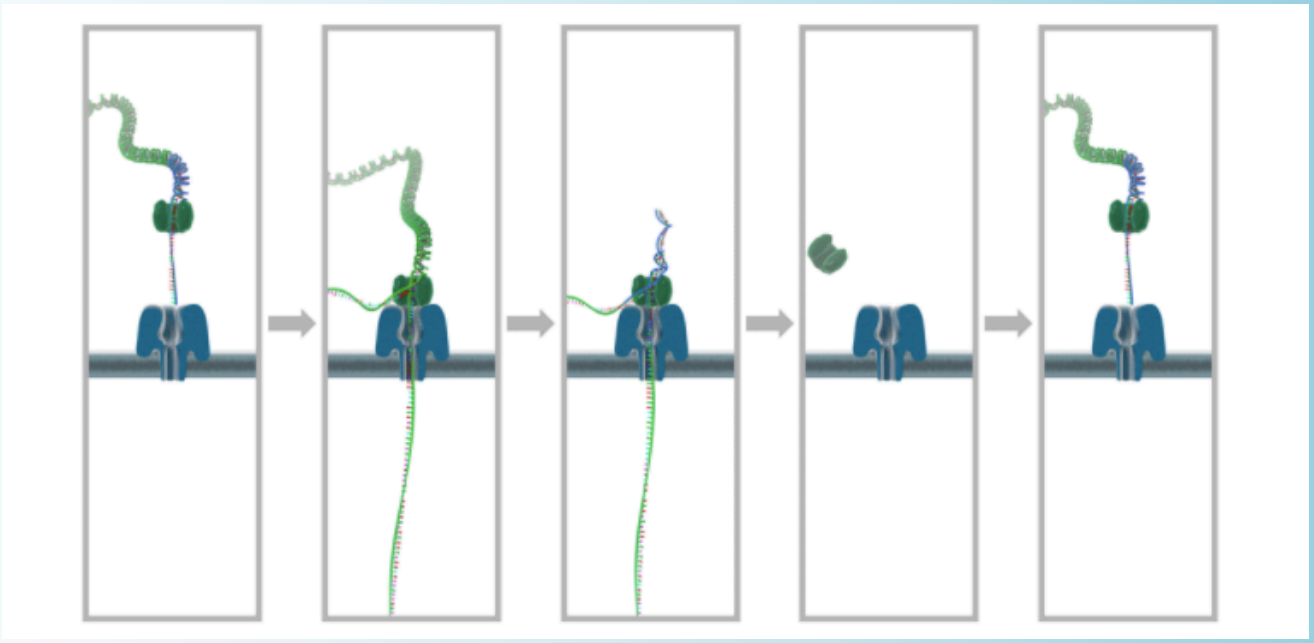
- Very high raw error rate (prone to random indel errors)
  - long PacBio reads that pass through the same insert multiple times reduce the consensus error rate to a more reasonable number
  - However, the number of sequencing passes and the consensus read length are a trade-off: longer sequences yield fewer passes and thus lower accuracy, and *vice versa*.
  - Can be corrected with less expensive Illumina short-read data
- Up to 10 times more expensive per gigabase than Illumina data
- Large insert libraries require a lot of DNA (20 µg +) that must be HMW and free of nicks and contaminants
  - PacBio libraries are single-molecule which requires more input material but also reduces biases

Superb fairywren genome (1.2Gb: 20x PacBio coverage)

- Required 80 µg of DNA per library prep (20 µg per library). **Best quality DNA is essential**
- This project used 27 runs on PacBio RSII: mean read length = 9000bp (very consistent: 7K-10K), mean number of reads = 99,000 (varied from 18K-122K), average of 870 megabases/run
  - **The Sequel (current instrument) produces 3-5 Gb per run and averages 8-12kb for large-insert libraries**
- About $10,000 total cost for this aspect of the project
- PacBio reads are often used in combination with Illumina for error correction and higher coverage. For this project:
  - 1 HiSeq 2500 lane of 100PE (17x coverage) & 1 MiSeq lane of 300PE (4x coverage) for genomic data and error correction (~$5000)
- Used other types of data for scaffolding: https://berkeley.box.com/v/Penalba-CBA-GenomeSequencing

Oxford Nanopore MinION

- Very inexpensive device for (potentially) ultralong reads
  - Large USB drive can be run off a laptop
  - Starter pack for DNA or RNA (1 device, 2 flowcells, 1 library kit) = $1000
  - Otherwise flowcells are $500-1000; libraries ~$100 each prep



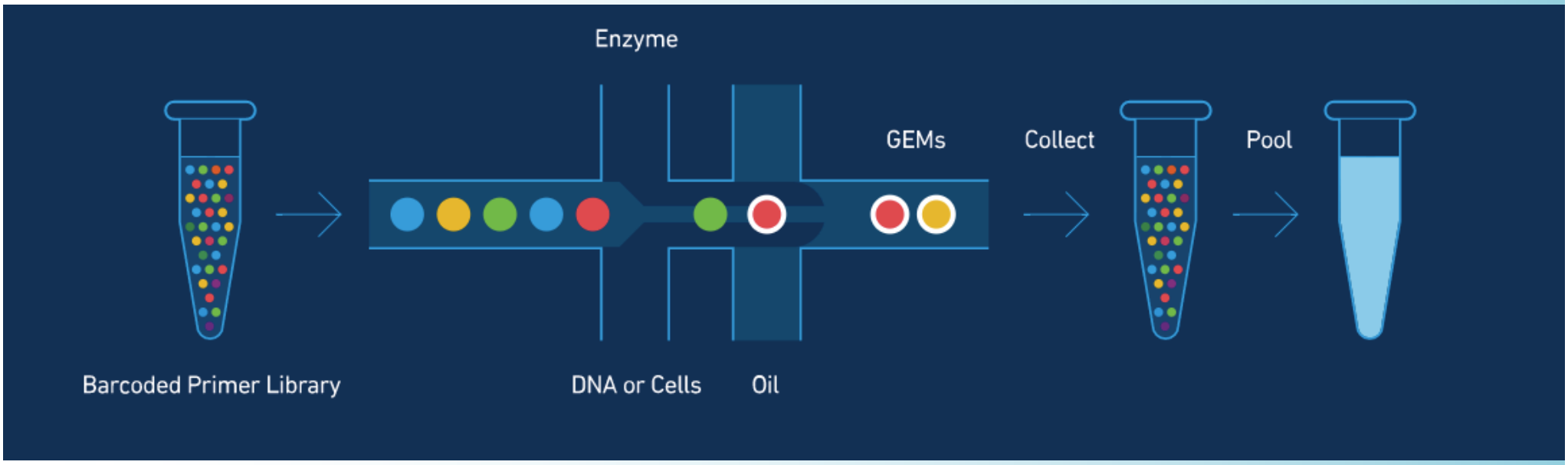
https://www.youtube.com/watch?v=GUb1TZvMWsw

Oxford Nanopore MinION

- Extremely DIY: currently best for researchers who enjoy experimenting
  - You make the libraries and run the instrument
  - Vendor (ONT) offers only limited support, slow to release data
  - On-line community and word-of-mouth are essential resources
- Unreliable with a high flowcell failure rate
- Requires very long, very high-purity DNA: any damage can truncate the read
  - Can use long-range PCR to try to repair damage
- Hypothetically 100kb+ per read but most current reads are very similar to PacBio
  - 2015 analysis https://doi.org/10.1016/j.bdq.2015.02.001
  - http://www.opiniomics.org/is-the-long-read-sequencing-war-already-over/
- However, new protocols and tweaks are always forthcoming
  - http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/
- Aaron Pomerantz (Berkeley) & Stefan Prost (Stanford)
  - http://www.thenextgenscientist.com/

10X Genomics Chromium

- Synthetic long reads: uses internal barcodes to bioinformatically link short-read sequencing from the same long fragment of DNA (100 kb!)
- Optimized for human genome so other organisms with different GC content may not work as well
- Available at UC Davis; Berkeley GSL can do HMW DNA extractions/assessments and organize sample transport.
- Berkeley has a similar instrument that barcodes single cells for RNA-Seq



Ugh, this sounds super-complicated. Can't someone else just do the work for me?

Berkeley has two terrific core facilities as part of the QB3 umbrella: GSL in Stanley and FGL in LSA: http://qb3.berkeley.edu/gsl/rates/

- DNA extractions
- Library preparations (Genomic DNA, RNA, 16S/ITS metagenetics)
- Quality control
- Captures

Great idea for people with very small sample sizes

- No need to learn molecular lab skills for a few one-off libraries; leave it to the experts

That said, getting to the point where you have the project design all planned out, know the key lab work choices, and understand how they will tie in with your bioinformatics approaches in order to answer your underlying scientific questions, well, that is the hard part

- Lab work can be kind of fun and can offer a new set of skills
- For researchers with the time to devote to it, you can save a lot of money
- Actually doing the process is the best way to learn it and to improve it

Shana McDevitt's key advice

GSL director: shana.mcdevitt@berkeley.edu

- Make sure you know what you will do with your analysis BEFORE you begin benchwork
- Indexes and insert sizes matter: make the wrong choices and you have backed yourself into a corner
- Understand base balance with respect to low-complexity libraries. Some library types require 10-25% PhiX to be successfully clustered (ddRAD, single amplicons)
- Plan ahead: don't assume that the GSL can help at the last-minute. (Shana gets 100+ e-mails per day!)
- Good quality libraries take time (4-8 weeks in the GSL)
- Contact her or GSL staff BEFORE submitting your first samples or libraries
- Carefully follow GSL submission guidelines to receive the best quality data
- The GSL can't help you make decisions about the biological questions (instead, consult with your PI, the CGRL, collaborators…)
- The GSL can't pre-bill for their work (so, again, plan ahead if funding is ending)

Bioinformatics

You will spend more time thinking about and looking at your data than collecting it, and for most of us, that's a good thing.

Berkeley CGRL: http://qb3.berkeley.edu/cgrl/

- https://github.com/CGRL-QB3-UCBerkeley

UC Davis: http://bioinformatics.ucdavis.edu/

- Training workshops: http://bioinformatics.ucdavis.edu/training/events/
- They estimate to budget as much for bioinformatics as for data collection when using their tearm of bioinfomaticians for analysis: http://bioinformatics.ucdavis.edu/services-2/

In Summary

- Project design & molecular work is the foundation for everything that follows: the best bioinformatics tools cannot salvage poor data
- Consult early; consult often
- Plan out as much of your project as possible before even starting extractions; when in doubt, choose the option that provides the most flexibility to protect yourself from unforeseen changes
- Budget is important but should not be the primary factor in study design; collecting the wrong data for your study questions will not lead to a successful project no matter how cheap
- High quality nucleic acid extractions are the key to high quality libraries which are the key to high quality data
- Take your time with the lab work; better to do it slowly and correctly than rush and waste time/money

Researcher Advice: Bioinformatics

- *Learn how to write code in perl/python as quick as possible and become familiar with computational practices and methods*
- *You will spend more time thinking about and looking at your data than collecting it, and for most of us, that's a good thing.*
  - So be sure that your data is good (high-quality and right for your analysis and questions.) since there is no fixing it once you get to this stage without $$$
- *I spend a lot of time on seqanswers and biostars. Also reading the manuals of particular programs are very helpful and some programs have 'vignettes' that you can follow with your own data which is also helpful.*
- *Find analysis tutorials on-line where you can learn the steps and then play around with a test data set*
- *Take advantage of free seminars (CGRL, on-line) to learn data analysis tools. Even if you end up using different programs for your own work, it will familiarize yourself with working in the command line and the key parameters you need to think about for making the right analysis choices with your data.*

Appendix I: High molecular weight DNA extraction resources

- http://stream.dcasf.com/webinar/olga-vinnere-petterson-dna-quality-requirements-for-single-molecule-sequencing/
- https://support.10xgenomics.com/genome-exome/index/doc/technical-note-sample-preparation-recommendations-for-the-chromium-genome-kit
- http://enseqlopedia.com/2017/03/hmw-dna-extraction-long-reads-nanopore-10xgenomics/

Appendix I: High molecular weight DNA extraction protocols & kits

Best methods: avoid columns, centrifugation, and pipetting of any DNA in solution (pouring is preferred)

- Qiagen MagAttract HMW DNA Kit kit or similar non-kit protocols with SPRI (doi 10.2144/000114460)
  - MagAttract used by UC Berkeley Nanopore researchers
- Phenol-Chloroform-Isoamyl extraction with SPRI clean-up of aqueous layer
  - Used in EGL for PacBio extractions with moderate results (~30 kb)
- GSL offers HMW extractions for $1000 for the first sample, $250 for re-extractions from the same tissue

Best tissue sources: avoid organs high in nucleases like liver and spleen. (Blood is handy since it digests easily; esp. good option with birds.)

Appendix II: Starting to think about de novo WGS

- https://www.molecularecologist.com/2017/03/whats-n50/
- http://omicsomics.blogspot.co.uk/2017/03/chromosome-scale-scaffolds-and-state-of.html
- https://blog.genohub.com/2017/06/16/pacbio-vs-oxford-nanopore-sequencing/
- http://cgrlucb.wikispaces.com/file/view/Assembly_Workshop_Berkeley_2017.pdf/618591217/Assembly_Workshop_Berkeley_2017.pdf
- https://berkeley.box.com/v/Penalba-CBA-GenomeSequencing